Proceedings of the 14th International Conference on Computational Creativity

Editors: Alison Pease · João Miguel Cunha · Maya Ackerman · Daniel G. Brown

June 19 — 23, 2023 · Waterloo in Ontario, Canada

Association for Computational Creativity

International Conference on Computational Creativity

ICCC'23

Proceedings of the Fourteenth International Conference on Computational Creativity

ICCC'23
Ontario, Canada — 19 - 23 June

Alison Pease, João Miguel Cunha, Maya Ackerman, and Daniel G. Brown (Editors)

Published by the Association for Computational Creativity (ACC)

Association for Computational Creativity

University of Waterloo
Canada

http://computationalcreativity.net/iccc23/

First published 2023

TITLE: Proceedings of the 14th International Conference on Computational Creativity

EDITORS: Alison Pease, João Miguel Cunha, Maya Ackerman, and Daniel G. Brown

ISBN: 978-989-54160-5-9

Published by the Association for Computational Creativity (ACC)

ISBN 978-989-54160-5-9

9 789895 416059

Preface

From June 19 to June 23, 2023, researchers from North America, Europe, Asia and Australia gathered at the University of Waterloo, in Waterloo, Ontario, Canada, for the fourteenth annual International Conference on Computational Creativity, ICCC. This was the first time that ICCC has visited Canada, and we were delighted to welcome our guests from so many different places, many of whom had never visited this country before. We offered a diverse collection of keynote presentations, all by Canadian scholars, and celebrated the bounty of Canadian produce, both in food and drink.

It is a challenge to run a small international conference that places a priority on bringing our community together. This year was no exception: visa delays, inflation in the cost of travel, and a number of community members who cannot travel due to concerns connected to COVID-19 all combined to mean that we delivered a lightweight hybrid conference so that remote participants could present their research, watch the technical sessions and participate in the Q+A periods in the research talks, while we still emphasized the experience of the onsite participants.

The conference was held at Federation Hall, on the University of Waterloo campus. Fed Hall was once the campus nightclub. One keynote speaker, Stacy Allison-Cassin, studied at Wilfrid Laurier University, another university in the city of Waterloo. She said she was amused to be presenting her research in a venue where she used to listen to concerts. We are grateful to the Fed Hall staff for their enthusiasm in hosting our conference, and in particular to the catering staff, who fed us very well. We managed the technical challenges via the able assistance of technical staff from Waterloo's Computer Science Computing Facility, who were excited to be part of hosting the first computer science conference at Waterloo in several years.

For the scientific program, ICCC'23 received 142 submissions: 61 to the long paper track and 81 to the short paper track. We set the deadline for the short paper track to May 2, just seven weeks before the conference, which meant that the presentations at ICCC were of research *à la minute*, which added to the excitement of the event. From the 61 long paper submissions, 22 were accepted as oral presentations. From the 81 short paper submissions, 21 were accepted and presented as short talks, 19 as posters and 10 as demos.

The accepted submissions have made ICCC'23 an internationally diverse venue, with work of re- searchers and practitioners from 62 academic institutions, spread over 20 countries.

As in past years, ICCC'23 recognized the excellency of three accepted submissions in different best paper categories:

- Best Paper: "Beyond Prompts: Exploring the Design Space of Mixed-Initiative Co-Creativity Systems" by Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, Mark Riedl
- Best Short Paper: "Diversity is Not a One-Way Street: Pilot Study on Ethical Interventions for Racial Bias in Text-to-Image Systems" by Kathleen Fraser, Svetlana Kiritchenko and Isar Nejadgholi
- Best Student Paper: "Interdisciplinary Methods in Computational Creativity: How Human Variables Shape Human-Inspired AI Research" by Nadia Ady and Faun Rice

In addition to the paper presentations, we had an exciting slate of other events as part of the scholarly program of ICCC.

On June 20, Rafael Pérez y Pérez launched the book he and Mike Sharples have recently published with Oxford University Press, *An Introduction to Narrative Generators*, to a very supportive audience of friends and colleagues.

We enjoyed three keynote speeches by three Canadian scholars:

- Jessica Thompson, Stratford School of Interaction Design and Business, University of Waterloo, spoke on June 20 on the topic, "Listening in place: How the social distribution of creative algorithms can help us understand cities."
- Stacy Allison-Cassin, School of Information, Dalhousie University, spoke on June 21 on the topic, "Absence versus Presence, Certainty versus Ambiguity: Creative Approaches to Ethical Data Practice."
- Kory Mathewson, Google DeepMind, spoke on June 22 on the topic, "The Show Must Go On: Co-Creating with Artificial Intelligence."

We had two panel discussions:

- On June 21, the Industry Panel discussed the theme, "Creative Machines Unleashed: Perspectives from Industry, Academia, Venture & Ethics", with Maya Ackerman as the moderator.
- On June 22, the Artists Panel discussed the theme, "On the Impact of Technology and AI on Artists' Practice and Careers", with Dan Ventura as the moderator.

We had a lively demo and poster session on June 22, made up exclusively of short papers submitted to the conference, which was organized by the Demo Chair Rob Saunders. The session was chaired by João Miguel Cunha and each presenter gave a 20-second flash teaser for their poster or demo, and participants enjoyed an exciting 90 minutes of scientific discussion.

ICCC also hosted one workshop, "Fictional Abstracts: Ethics, Sustainability and Creative-AI Futures", brought to us by Petra Jääskeläinen and Camilo Sanchez, on June 19. Also on June 19, Tony Veale presented a tutorial, "Why so serious? Building creative systems with a sense of humour."

Finally, on June 20, ICCC hosted a doctoral consortium, where 15 graduate students were able to present their scholarship to their peers and to receive feedback from senior academic mentors from our community. This event was chaired by Pablo Gervás.

Social program

ICCC is a small, but vibrant community, and our vibrancy is enabled by bringing us all together at a variety of social events during the conference. This year, we ran two specific social events: a reception at the Delta Hotel after Jessica Thompson's keynote speech on June 20; and an excursion to the Niagara Peninsula on June 23, featuring a visit to Chateau des Charmes Winery in Niagara-on-the-Lake for a winery lunch, and a few hours enjoying the spray of the waterfalls of Niagara Falls.

Acknowledgments

The ICCC 2023 organizational team expresses our gratitude to our sponsors: the University of Waterloo's Faculty of Mathematics, who also offered organizational assistance to the conference; its Faculty of Arts (who supported the fees for the artists who participated in the panel on June 22); its Library (who supported arrangements for Prof. Allison-Cassin's keynote); and the David R. Cheriton School of Computer Science at the University of Waterloo. In particular, we are grateful to the staff of the Cheriton School's Computer Science Computing Facility for their technical support of the conference. The Association for Computational Creativity (ACC) has offered institutional and organizational support, and ensures the high scientific and creative standards of our community.

We also appreciate the Contemporary Art Forum, Kitchener and Area (CAFKA), with whom we partnered in organizing the artists panel. This event was part of CAFKA's 2023 biennale, *Stay With Me*.

We appreciate the guidance of ACC's steering committee and the labour of the program committee and organizational team, the external reviewers, and the other local staff who made the event possible. Finally, we thank you, the reader: we hope the content of these proceedings interests you, and look forward to what research and other creative contributions will be informed by the research from ICCC 2023!

Organizing Committee

General Chair
Dan Brown

Program Chairs
Alison Pease
João Miguel Cunha
Maya Ackerman

Workshops and Tutorials Chair
Rafael Pérez y Pérez

Doctoral Consortium Chair
Pablo Gervás

Demo Chair
Rob Saunders

Publicity Chair
Christian Guckelsberger

Proceedings Chair and Webmaster
Paul Bodily

Program Committee

Senior PC Members

Alison Pease – University of Dundee
Amílcar Cardoso – University of Coimbra
Andrés Gómez de Silva Garza – Instituto Tecnológico Autónomo de México
Anna Jordanous – University of Kent
Ashok Goel – Georgia Institute of Technology
Carlos León – Universidad Complutense de Madrid
Dan Ventura – Brigham Young University
Daniel G. Brown – University of Waterloo
François Pachet – Spotify
Geraint Wiggins – Vrije Universiteit Brussel / Queen Mary University of London
Hannu Toivonen – University of Helsinki
Julian Togelius – New York University
Kazjon Grace – University of Sydney
Maria M. Hedblom – Jönköping University
Mary Lou Maher – University of North Carolina at Charlotte
Nada Lavrač – Jožef Stefan Institute
Oliver Kutz – Free University of Bozen-Bolzano
Pablo Gervás – Universidad Complutense de Madrid
Penousal Machado – University of Coimbra
Rafael Pérez y Pérez – Universidad Autónoma Metropolitana at Cuajimalpa
Rob Saunders – Leiden University
Simon Colton – Queen Mary University of London
Tony Veale – University College Dublin

PC Members

Alan Tapscott – Universitat Pompeu Fabra
Allison Parrish – New York University
Amy K. Hoover – New Jersey Institute of Technology
Amy Smith – Queen Mary University of London
Ana Rodrigues – University of Coimbra
Antonios Liapis – University of Malta
Bob Sturm – KTH Stockholm

Brad Spendlove – Brigham Young University
Brian Magerko – Georgia Institute of Technology
Caterina Moruzzi – University of Konstanz
Christian Guckelsberger – Aalto University
Colin G. Johnson – University of Nottingham
Constant Bossou – Santa Clara University
David C. Brown – CS Dept. WPI
Diarmuid P. O'Donoghue – Maynooth University
Doug Van Nort – York University
Enric Plaza – IIIA-CSIC
Evana Gizzi – Tufts University
Georgios N. Yannakakis – University of Malta
Giovanni Pilato – ICAR-CNR
Guendalina Righetti – Free University of Bolzano
H. Sofia Pinto – IST, University of Lisbon
Heather Ligler – Pennsylvania State University
Hugo Gonçalo Oliveira – University of Coimbra
Ivan Guerrero Roman – UNAM-IIMAS
Jer Hayes – Accenture
João Correia – University of Coimbra
João Miguel Cunha – University of Coimbra
Jon McCormack – Monash University
Josep Blat – Universitat Pompeu Fabra
Kıvanç Tatar – Chalmers University of Technology
Lakshmi Nair – Georgia Institute of Technology
Liuqing Chen – Zhejiang University
Lonce Wyse – National University of Singapore
Luis Fabricio Wanderley Goes – University of Leicester
Margareta Ackerman – Santa Clara University
Maria Teresa Llano Rodriguez – Monash University
Martin Thiering – Technische University Berlin
Martin Znidarsic – Jožef Stefan Institute
Matthew Guzdial – University of Alberta
Max Kreminski – Santa Clara University
Maximos Kaliakatsos-Papakostas – Aristotle University of Thessaloniki
Melissa Roemmele – Language Weaver (RWS Group)
Michael Cook – Queen Mary University of London
Mihailo Antovic – University of Nis
Milena Fisher – The Creativity Post
Najma Mathema – Brigham Young University

Nuno Lourenço – University of Coimbra
Oliver Bown – Interactive Media Lab, UNSW
Paul Bodily – Idaho State University
Pedro Martins – University of Coimbra
Philipp Wicke – Ludwig Maximilian University
Philippe Pasquier – Simon Fraser University
Piotr Mirowski – DeepMind
Prashanth Thattai Ravikumar – De Montfort University
Rafael Peñaloza – University of Milano-Bicocca
Roberto Confalonieri – Free University of Bozen-Bolzano
Roisin Loughran – Dundalk Institute of Technology
Santiago Negrete-Yankelevich – Universidad Autónoma Metropolitana
Sarah Harmon – Bowdoin College
Senja Pollak – Jožef Stefan Institute
Shlomo Dubnov – University of California in San Diego
Simo Linkola – University of Helsinki
Taisuke Akimoto – COMOS
Tapio Takala – Aalto University
Tarek Richard Besold – TU Eindhoven
Thomas Winters – KU Leuven
Tiago Martins – University of Coimbra
Tomi Männistö – University of Helsinki
Wendy Aguilar – UNAM
Ying-Ping Chen – National Yang Ming Chiao Tung University

Contents

1. Language and Storytelling

Have I Got Views For You!
Generating “Fair and Balanced” Interventions into Online Debates

Tony Veale
School of Computer Science
University College Dublin
Belfield, Dublin D4, Ireland
Tony.Veale@UCD.ie

Abstract

Online debate around divisive topics has become increasingly fractured, leading to the emergence of “echo chambers” in which disputants communicate almost exclusively with those who hold compatible views. To inhibit the growth of echo chambers and expose disputants to both sides of an argument – in ways that encourage dialogue across the divide – we aim to automate the generation of creative interventions into otherwise insular online debates. On highly echoic platforms such as Twitter, bot-driven interventions run contrary to best practices, and may be reported as an abuse of the system. However, passive interventions can instead use story generation to dramatise an ongoing debate. If the stories so generated are engaging and balanced, and are aptly labeled with attested hashtags, they can draw users to a bot's content, thus avoiding any need for a bot to elbow its content into a live conversation. The *Excelsior* system, as described here, aims for amusing, even-handed engagement by packaging its data-driven stories as comic strips which integrate two sides of any argument into a single visual intervention.

Hold The Funny Pages!

It has been suggested that life is a tragedy for those who feel, and a comedy for those who think. We see this dichotomy writ large every day on social platforms such as Twitter, where discourse around contentious topics generates an excess of polarizing feeling and a comparable dearth of rational thought. Such platforms incentivize the articulation of short, pithy positions that prize outrage over insight, and in which interactions between opposing camps fall quickly to rancour. However, even rancorous exchanges may be preferable to the non-engagement with antagonistic stances that is too often observed on Twitter, for at least they can expose users to multiple points of view. Instead, inward-looking, defensive structures called *echo chambers* (Barberá et al. 2015) insulate disputants from interactions with those with whom they are in dispute, and feed the growth of factionalism and the decline of real debate on Twitter.

Bots are an oft-aligned presence on Twitter, but one benign use of Twitterbots is the generation of interventions to foster engagement between holders of opposing views (Blaya 2019). Such interventions can cut to the heart of a dispute, by repackaging the nub of a conflict in an engaging form. Although many users follow bots out of an appreciation for their whimsical and oddly human-like outputs, few welcome unsolicited intrusions from bots in the form of direct messages, replies or mentions. Even bots that point out spelling errors may provoke vitriol or scorn (Veale and Cook 2017). After all, few of us like to be lectured by strangers, least of all automated strangers. Our goal in the *Excelsior* system is the generation of narrative interventions that are as engaging as they are unthreatening, and which users can find for themselves via the use of attested hashtags.

Key to this engagement is the use of comic strips as a narrative medium. These strips originate in the ”funny pages” of newspapers, where they were meant to entertain more than to educate, yet comics are a sequential art form (Eisner 1985; McCloud 1993) that is not limited to tales of funny animals and masked heroes. Here we aim for education *and* entertainment, to give data-driven stories about serious topics a harmless comedic form that is more likely to foster engagement than suspicion and outrage. Crucially, each newly created comic must balance two points of view, an argument and its converse, as articulated in the underlying data, which in the current case is the ongoing debate on Twitter about climate change, or vaccines, or guns, or abortion.

Excelsior proceeds by first identifying hashtags that convey a clear stance to a topic, such as *#EcoLiteracy*, *#FireFauci* or *#GetVaccinatedNow*, and then arranges related tags into sequences of mounting emotion, such as from curiosity to skepticism to disgust. An emotional inversion is performed mid-sequence, such as from disgust to admiration, to shift the narrative to an opposing viewpoint. The full sequence is then rendered as a comic, one panel per hashtag, that balances both points of view. The resulting comic can then be tweeted as an animated GIF along with the tags that punctuate its plot. Excelsior's approach to data storification does not aim to summarize the totality of a debate all at once. Rather, as we will show, it treats each debate as a space of views, and samples stories from this space in a way that, over time, cumulatively mirrors its emphases.

Back Issues: Related Work and Ideas

Comics are a medium for story-telling that requires a narrative impetus. For the *Comic Chat* system of (Kurlander, Skelly, and Salesin 1996), this impetus comes from the interactions of the users of online chatrooms. User texts are not summarized but placed verbatim into speech balloons above cartoon depictions of each user. Each conversational beat produces a single panel, and sentiment analysis is used to determine which variant of a user's comic avatar is associated with each speech act. But this impetus can also be machine-generated, and comics offer a viable medium for rendering automated stories, as in the story-to-comic generators of (Alves et al. 2007), (Pérez y Pérez, Morales, and Rodríguez 2012) and (Veale 2022). This can be modeled as a text-to-text generation task if each comic is specified using XML, as in the CSDL (Comic Strip Description Language) of (Alves et al. 2007), the CBML (Comic Book Markup Language) of (Walsh 2012) or the ComiXML of (Veale 2022). *Excelsior* builds upon the latter, ComiXML, as it allows a comic to be specified as a specific arrangement of visual assets, drawing from a repertoire of hundreds of different character poses and panel backgrounds.

This is a symbolic, componential approach to building comic strips, in contrast to the neural approaches typified by (Melistas et al. 2021) and (Proven-Bessel, Zhao, and Chen 2021). Neural approaches are trainable, and so are adaptable to specific data sets and visual genres (e.g., *Manga* in (Melistas et al. 2021), *Dilbert* in (Proven-Bessel, Zhao, and Chen 2021)). They are, in principle, capable of generating diverse images to match a given text prompt, although the visual outputs of the generative adversarial networks in (Melistas et al. 2021; Proven-Bessel, Zhao, and Chen 2021) are often blurry and ill-formed. Moreover, the relationship between image and dialogue, which is the crux of the comics medium, is difficult to control in such models. This relationship is crucial when comics are used to package interventions into a debate, especially when the goal is to balance opposing points of view.

Alternatively, images and text may be generated separately, by models that specialize in each. For instance, very large language models such as *GPT3* and *ChatGPT* can be used to generate stories for a given prompt (Xie, Cohn, and Lau 2023), in the desired form (e.g., a two-person dialogue, a one-act play). To provide a suitable context to the generator, the prompt may in turn be generated by existing narrative extraction methods (Santana et al. 2023), as applied to a debate corpus of interest. Individual text fragments can then be used to prompt an image generator such as *Dall-E* or *Stable Diffusion* (Gozalo-Brizuela and Garrido-Merchan 2023) to create a panel setting for each. But large language models (LLMs) are resource-intensive blackboxes that are not conducive to the development of small-footprint systems. Neither do LLMs yet permit easy interrogation of their logical processes, or offer guarantees as to whether their outputs convey the intended meanings. In contrast, a symbolic model can tick all of these boxes.

Data Collection, Organization and Analysis

We initially viewed each of the four debate spaces – climate change, vaccines, guns and abortion – as distinct, and collected four separate corpora of tweets via Twitter's streaming API, guided by seed sets of topic-related hashtags. We have come to realize that all four instantiate a single overarching debate concerning the acceptable balance of power between the state and the individual, and although each corpus has unique hashtags of its own, many tags – especially those of a political nature – recur across debate boundaries. Table 1 reports the number of distinct tweets and users comprising each corpus, noting how many are in fact retweets.

Dataset	# Tweets	# Retweets	# Users
Vaccines	1,624,173	1,244,009	391,489
Climate Change	1,017,087	691,333	340,836
Abortion	369,914	237,139	159,196
Gun control	205,535	131,728	62,387

Table 1: Size and makeup of the four debate datasets.

Table 2 reports the number of distinct hashtags in each dataset. While the raw counts (*# Tags*) are large, the number of distinct tags that convey a clear stance toward an explicit topic (*# Stanced*) is much smaller. These tags, in turn, conform to a smaller set of semantic *patterns*. These patterns are templates with semantic filler types that allow Excelsior to determine the stance and topic of each tag. For instance, the hashtag pattern #Fire{personal} is instantiated in 11 ways across the four debates, where {personal} can range from Fauci to DeSantis to Trudeau. The most varied patterns include #Get{solution} (30 fillers) and #No{solution} (29 fillers), #Pro{solution} (20) and #Anti{solution} (24), #Arrest{personal} (25) and #LetsGo{personal} (11), #Boycott{business} (12) and #Boycott{place} (20), #No{problem} (19) and #Stop{problem} (31).

Dataset	# Tags	# Stanced	# Patterns
Vaccines	39,366	4,236	1,986
Climate Change	32,375	1,993	1,043
Abortion	18,563	1,344	652
Gun control	16,090	957	470
All four debates	90,323	6,982	2,985

Table 2: Raw and processed hashtag counts per dataset.

Just as sets of domain-specific "seed" hashtags are used to collect each individual debate dataset via Twitter's streaming API, a set of seed entities is also used to drive the mapping of newly collected tags to generic patterns; e.g., "Greta"={person}, "covid"={problem} and "vax"={solution}. A bootstrapping process is used to identify candidate patterns among hashtags in which camelcasing indicates a multi-word structure, and for which sentiment analysis indicates a positive or negative stance, such as #LetsGoBiden, #PleaseVaxUp and #EndCovidNow. Replacing any known entities in these tags gives us the candidate patterns #LetsGo{person}, #Please{solution}Up and #End{problem}Now. These candidates are manually curated, and added to Excelsior's lexicon only when they convey a clear stance toward the referenced entity. But these additions can, in turn, be used to suggest new entities, by matching the pattern against other hashtags. For example, since #LetsGo{person} also matches #LetsGoDeSantis, the entity "DeSantis"={person} is also offered as an addition to the lexicon. These new entities then allow further patterns to be identified in the data, such as #{personal}2024, #LockUp{personal} and #{personal}Lies. A candidate pattern may unite multiple entities, such as #{personal}Failed{place}, which matches #TrumpFailedAmerica as well as #DeSantisFailedFlorida.

Each curated pattern is associated with a firm stance, either $accepting$ or $rejecting$, toward a referenced entity. This must be done manually because online debate is fast-moving and sentiment analysis is so often wrong. For example, #LetsGo{person} is actually a $rejecting$ rebuke, and not an $accepting$ endorsement of {person}, due the peculiar origins of the jeer #LetsGoBrandon. Each tag pattern is also linked to an emotional framing, which offers a finer view of the feeling being articulated. So #{personal}2024 evokes an $election$ framing while #LockUp{personal} evokes a $prison$ framing. A framing allows a hashtag to be visualized as a comic panel with apt character poses, apt dialogue and an apt backdrop. Thus, the $prison$ frame suggests someone holding keys to an other's cell, while the $election$ frame suggests one voting for another in a poll centre, and so on.

Each hashtag pattern is linked to one or more of 96 framings, such as $battle$, $freedom$, $contempt$ and $hoax$. A framing often represents a metaphorical perspective, such as $battle$ or $slavery$, or an intense feeling, such that a given problem or solution is a $hoax$. We choose them for their dramatic potential, as well as for their suitability to the sampled tags. Each dramatic frame is associated with a set of apt dialogue patterns, for both a protagonist (the main speaker) and an antagonist (one holding an opposing view). For instance, #Fake{solution}, #Phony{solution}, #{solution}Cult and #{solution}Con typify a large family of similar tags that are linked by the $hoax$ framing. In this context, the dialogue patterns "Expose {solution} as a fake!", "Unmask {solution} as a fraud!", and "I hate the hypocrisy of {solution}!" are available to the protagonist, while the patterns "What's the issue with {solution}?", "Why are you down on {solution}?" and "What's so wrong with {solution}?" are possible responses for the antagonist.

The set of 96 framings is organized as a graph. One framing links to another if the second adds to the feelings of the first, thus serving to build the debate (or comic) toward an emotional crescendo. So, for instance, $scepticism$ can lead to $denial$ or $blame$. which can lead to a call for $defunding$ or an accusation of $tyranny$ or $hoax$. In turn, $tyranny$ can lead to cries of $treason$ or $fascism$, where $treason$ can lead to calls for $prison$.

The ABCs of Comic Generation

This graph allows Excelsior to organize hashtags in the data into plot-like sequences that build to a dramatic climax, even if those tags were never used in the same tweets or even by the same users in the original dataset. Every random walk in this graph produces a valid plot, although Excelsior must then ground the constituent frames in actual hashtags that refer to the same topic. It is also not enough to articulate just one viewpoint on a topic. Rather, the "plot" should switch from one side to another at some turning point in the narrative, and thereby allow the antagonist to become the protagonist. To facilitate this switch of perspectives, the graph also links framings to those that express opposing emotions. For instance, $treason$ is thus linked, by opposition, to $heroism$, $election$, and $admiration$. Note however that these transitions at the framing level are only pursued if there are actual hashtags in the data to support them. A plot can switch from $treason$ to $election$ with regard to topic X only if the data contains tags that imply that X is a traitor, and tags that call for X to be elected.

To dictate the general shape of a plot, we employ the AAB string notation. The place holders A and B can denote any framing, but the sequences AA and BB can only denote a transition from one framing to another more intense framing on the same side of the debate, as allowed for by the framing graph. Conversely, AB and BA can only denote a transition between frames on either side of the debate, as allowed for by the graph.

A plot with the shape AAB is thus realized as a comic in which a particular stance toward a given topic is established in one frame/panel, intensified in the second, and rebutted in the third. This generic AAB pattern is an example of what (Loewenstein, Raghunathan, and Heath 2011) call a repetition-break structure, in which a norm is first established by repetition and then dashed to produce a humorous or creative effect. Those authors provide evidence for the pattern's popularity and effectiveness in eye-catching TV adverts, while (Loewenstein 2018) argues for the utility of the pattern in constructing materials designed to spread rapidly across social networks. We further generalize the AAB pattern here to allow for controlled repetition of the norm and its opposite. Fig. 1 presents a comic created by Excelsior for the pattern $AAAABBB$, as applied to the joint dataset. The system picks the topic $climate\ change$, and balances views for and against the topic in the comic.

The joint dataset combines tweets and tags from all four of the debates in Table 1. When an explicit topic is provided, such as $carbon$, Excelsior confines itself to tags that focus on that topic. To offer the data-fitting process some wiggle-room, we define a topic graph to connect related ideas for which a stance toward one translates to a stance toward the other, such as $climate$ and the $environment$, $carbon$ and oil, or $Biden$ and the $Democrats$. This allows Excelsior to veer from one topic to another when instantiating its AABs, to generate more varied comics while staying on-message.

As shown in Fig. 1, each hashtag that instantiates the A/B elements of the $AAAABBB$ pattern is given its own panel, under which the original tag is displayed. Each comic uses two characters, which are rendered in blue and red to make them visually separable. This visual identity is important when the viewpoint switches from one side of the debate to the other, as happens here in the second panel of row two. The protagonist, shown in blue, advances the A side of the argument on climate change, and here advances the pro-green agenda. The antagonist, shown in red, responds with as many questions as rebuttals. Excelsior strives for balance across panels and within panels too, and generally aims to let no claim go unquestioned, whatever its validity. When the agonists switch sides, it becomes red's turn to voice the anti-green B side in the face of blue's advocacy.

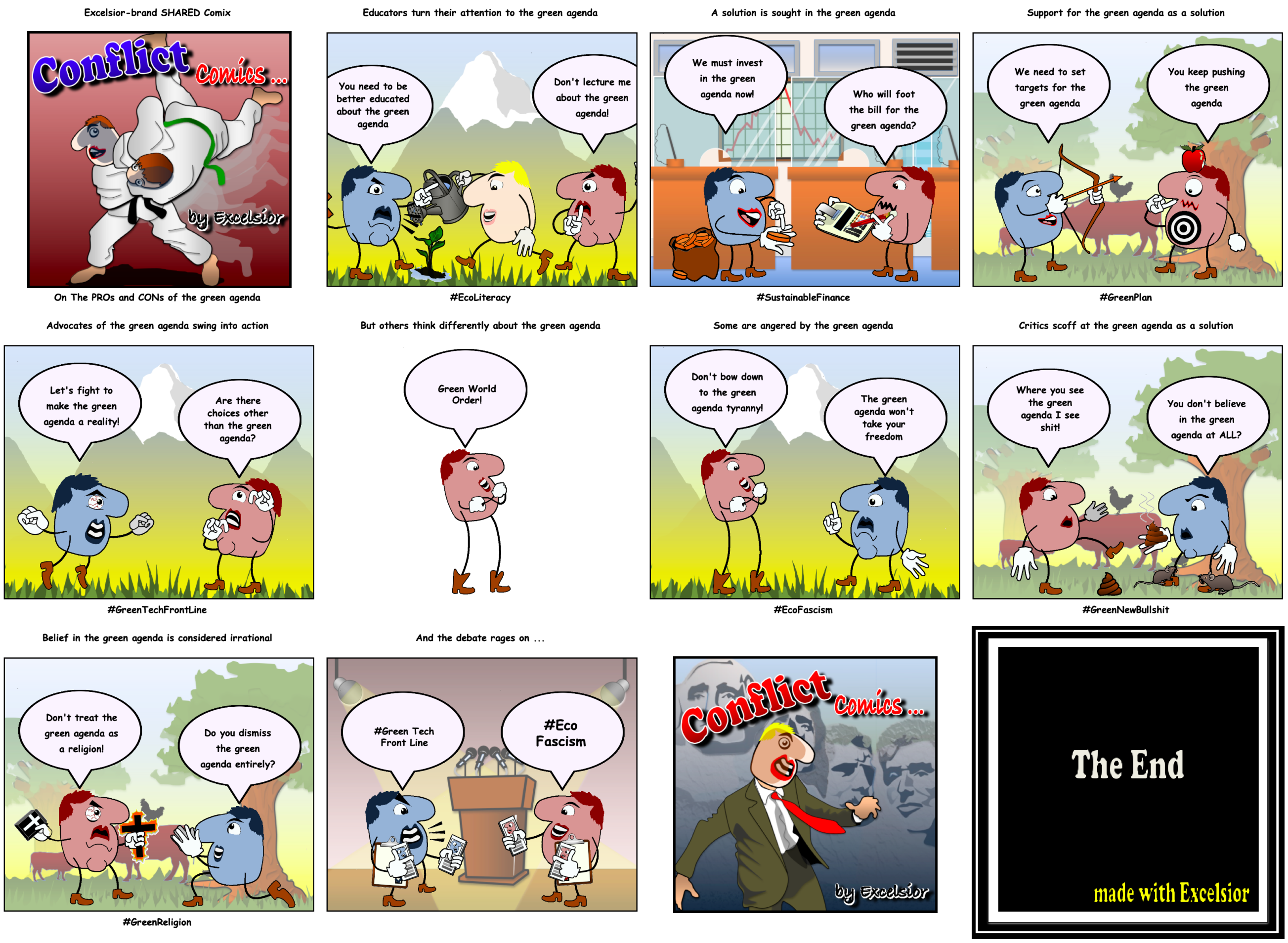


Figure 1: An Excelsior comic in the domain of climate change. Stance reversal occurs in the 2nd panel of row 2.

The AABs of Irony Generation

(Rozin et al. 2006) show that the AAB pattern is more effective than any other variation (e.g., AB or $AAAB$) at inducing a humorous response to a creative stimulus. One consequence of using comic strips to package the products of data "storification" is that stances which are already emotionally intense are tipped into humorous exaggeration by a vividly expressive rendering. The AAB pattern is used here to inject conflict and balance into each comic, but any emergent humour is ultimately unplanned. Still, we can foster humour by using the AAB pattern in its purest form, with data that has been explicitly chosen for its humorous potential.

The internet is replete with humorous content, such as joke lists, that can be injected into a comic. These resources, though large, are often problematic, since they lean heavily on racism, sexism and homophobia. (Tang et al. 2022) present a transformer for detecting offense in Reddit joke lists, but offer no means of controlling the meaning of a joke or making it fit a given context. There is little point in forcing an arbitrary joke about farming, say, into a comic on this topic if Excelsior cannot know which side of a debate the joke is on. Rather, we need a more controlled source of humour that cleanly interfaces with the assertions implied by each hashtag. For this we turn to the "about" similes of (Veale 2012).

Humour involves playful insincerity, so to avoid serious misunderstandings, humorists often provide subtle but predictable cues to their insincerity. In the case of exaggerated or ironic similes, these cues are found in hedge words like "about" or "almost." Take the heavily panned film *Cats* (2019). After viewing an unappealing trailer, one might describe the film as "*about* as marketable as a flesh-eating virus." These cues serve a dual function: they signal a creative intention on the part of a writer, and allow machines to trawl large quantities of creative similes from the web. (Veale 2012) reports that such a trawl pulls in a large set of ironic similes, in which one quality is asserted but its opposite is implied, and a larger set of comical similes whose qualities are asserted literally. If Excelsior can infer the qualities implied by a specific tag framing, it can make the qualities comically explicit by using vivid similes from this corpus. It can also exploit the AAB pattern to magnify the humour of its choices, by chasing two literal similes (AA) for an implied quality with an ironic twist (B).

Figure 2: An Excelsior comic on the topic of vaccination which follows an AAB irony pattern.

(Hao and Veale 2010) present a means of separating ironic from literal "about" similes, noting that positive qualities (like $marketable$) are often intended ironically, while negative ones are more often intended literally. As noted earlier, Excelsior maps hashtags like #FireFauci to patterns such as #Fire{person}, and further maps those patterns to framings like $rejection$ and $contempt$. We now associate these framings with the qualities they imply of their referents, for instance, that the referent of #Fire{person} is neither competent nor welcome, or that the focus of #{solution}Farce is hardly credible. An AAB pattern can now be crafted from a single tag like #JabFarce, as illustrated in the comic of Fig. 2. Note how the quality $credible$ is treated literally for two comparisons before it is subverted by irony in a third. Irony offers balance even in the case of a single hashtag.

Nonetheless, Excelsior is careful to balance the scales. Just as the comic opens with a panel visualizing the tag #JabFarce via the framing $contempt$, it closes with one visualizing an opposing view, #AntiVaccineMadness, via the antithetical framing $defence$. The core conflict between these views is then summarized in a final panel.

Experiments in Transformation

As a generator of topical comics, the Excelsior system is both knowledge-driven and data-driven. Its comics reflect real tensions in social-media data sets that are growing and evolving in real-time, and it uses top-down knowledge-structures to make sense of this data. The comics themselves are specified using an XML schema that assembles a fixed repertoire of poses and settings into LEGO-like dioramas, but they are filled with dialogue that, while apt, relies on prescripted templates. These trade-offs make Excelsior responsive *and* controllable, but the surprises in its comics come from the data, which is constantly changing, and not the system's own knowledge, which evolves at a much slower pace.

Symbolic systems make poor learners, but they can still serve as good teachers. To see why, consider how a pretrained neural model is fine-tuned for a new task. A transformer language model such as the $T5$ (Raffel et al. 2020) can be further trained on a set of input/output text pairs, so that it can learn to map from a given input text to the desired output text. We can, for instance, fine-tune a $T5$ on a set that maps domain-specific tweets onto the XML specifications

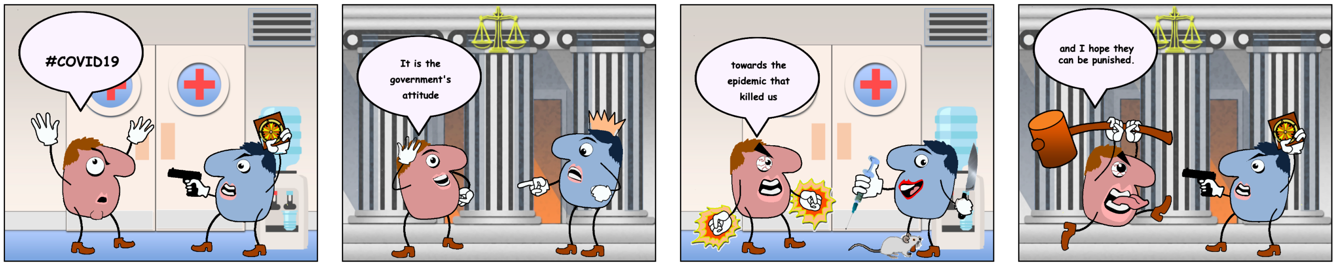


Figure 3: A comic generated by a T5-small transformer that is fine-tuned on a dataset of Covid/vaccine tweets.

of the corresponding comics. Indeed, a $T5$-$small$ model of 60 million parameters is sufficient to the task of learning the text-to-XML mapping for a domain such as vaccines or climate change. We can see this with a corpus of 1,500 Covid/vaccine tweets which have been manually annotated with ComiXML; of these, we hold back 150 for validation and 150 for testing. Fig. 3 shows the comic that is rendered from the XML output for an unseen test tweet: "#Covid19 It is the government's attitude towards the pandemic that killed us and I hope they can be punished." Note how the dialogue is one-sided, and repurposes the text of the tweet, but does so in a way that is visually expressive and emotionally apt.

Notice also how the action switches from a hospital setting (where the dialogue touches on medical issues) to a legal setting (where the dialogue touches on governance and law) and back again. Because its fine-tuning tweets are segmented by XML mark-up tags, the transformer learns to segment each new tweet into balloon-sized morsels of dialogue. In each case, the transformer assigns poses to the agonists that match both their explicit interactions (e.g., panels 2-4) and their implicit stances (e.g. panel 1). Here the opening panel aptly sets the scene, and serves to foreshadow the scepticism of the protagonist (in red) in the following panels.

The $T5$ performs well on new tweets, and learns how to use the ComiXML schema well in the Covid/vaccine domain. The value of XML as an output format cannot be understated, as it allows a generator to automatically check the validity of the transformer's outputs. On the rare occasions when these are malformed – e.g., when the XML is not schema-compliant, or when it invents new poses or settings – a new output can be re-sampled from the same input.

But the transformer does not generalize well beyond its specific domain. When presented with tweets lacking an overt focus on Covid or vaccines, it cannot but view them through a monocultural lens. It continues to place characters in hospital and graveyard settings, as though perceiving a subtext that is invisible to human readers. When we repeat our experiments with a new fine-tuning corpus of 1500 tweets, this time on the topic of climate change, we observe the same outcome. The transformer performs very well on new in-domain tweets, but does not generalize robustly beyond this domain. The Covid transformer is well-formed but inept in its handling of climate issues, and the climate transformer is similarly inept in its handling of vaccines. The situation improves when a transformer is fine-tuned on a joint corpus for both domains, but it still fails to generalize well to other domains, such as gun control and abortion. Moreover, it is costly to fine-tune the $T5$ for each new domain. We find that it takes 2 to 3 person weeks of effort to collect and mark up each new set of 1,500 training tweets.

This is where a symbolic teacher can step in. Though its dialogue patterns are limited in number, such a teacher can generate dialogue for specific topics in a new domain. Its outputs will be guided by attested hashtags in the domain, so it will produce short texts that are representative of the feelings swirling about those topics in the given dataset. It can also produce the XML comic specifications for those texts, to automatically generate both sides of the input/output training pairs for the transformer. A symbolic Excelsior can thus lend its ability to generalize, via templates, to a learner with an unsure footing in a new domain. The $T5$ can now be periodically fine-tuned on the new example sets.

Template-based generation becomes more stilted and predictable with time. Excelsior's dialogue model can talk about new topics, but only in the same old ways. To expose a learner like our $T5$ to fresh ideas and fresh ways of talking, we need fresh data. Fortunately, a symbolic teacher that can interpret new hashtags in terms of existing patterns can easily find tweets that use those tags. It can fine-tune the learner by pairing these tweets with comics it produces from the tags. in this way, a symbolic teacher can greatly reduce the time taken to create a training set for a new domain.

We have some way to go before the symbolic Excelsior is inevitably usurped by its statistical student. For now, only the symbolic model can offer a complete solution to the generation of comic strips that balance the views of multiple users across competing "echo chambers." This model will be replaced piecemeal rather than all at once, as transformers learn to improve on its individual parts.

Moral Dimensions and Dilemmas

The generation of comics with carefully balanced meanings is a means to an end rather than an end unto itself. These comics serve as interventions into a fractious online debate, so as to expose disputants to all sides of an issue. They are not intended to provide answers but to raise questions and foster discussion. Yet, in doing so, they also pose some difficult questions for their creators.

Some disputes make it difficult to stay above the fray. Is there a moral imperative to take a side when some actors spread conspiracy-fuelled misinformation and play fast and loose with scientific facts? Balance is surely a desirable quality, but is it always right or wise to give exposure to extreme views in the interests of fairness? Each time we encourage debate between opposing sides, we run the risk that more, not fewer, people will adopt the controversial views that we put under the spotlight. Yet, to serve as an honest broker that appeals equally to both sides, a creative system cannot afford to be partisan. This refusal to hold opinions of its own can make a creative system seem indifferent and amoral, a purveyor of what (Frankfurt 1986) famously called "bullshit." It is, it seems, a question of balancing one harm against another: are echo chambers so detrimental to our social discourse that these other risks are worth taking?

A "fair and balanced" creative system can manifest bias in subtle ways. For instance, it might always grant the last word on a topic to one side of a debate, e.g., to show there is a clear reply to every objection to vaccines or to every doubt about climate change. The ordering of claims in a comic can make a certain position seem like an argument's end-point rather than a starting point. A system that uses humour to promote engagement may not use its humour even-handedly, and may, for example, make certain views the preferred butt of its jokes, or use more risible visual representations of those views. We must give systems knowledge but not opinions, and be shrewd enough to distinguish one from the other. This is challenging whether one is building a top-down symbolic system or a bottom-up statistical system.

The most trenchant views on Twitter often involve *ad hominem* attacks, but should a system repeat these even if it balances them with supportive counter-points? In politics, such attacks are a way of life and the cost of doing business, but what of others in the public sphere? Public figures make for good "extras" in a comic strip, because they lend an emotive face to non-visual ideas. The white-haired figure of Tony Fauci and the pig-tailed figure of Greta Thunberg bring concepts such as public health policy and climate change down to human-scale. Since each is an effective champion *and* a lightning rod for controversy, Excelsior uses both in its comics, for the same reason they anchor so many tags in the data. We want Excelsior to treat all public figures equally, but do not want to aid the demonization of certain individuals. Excelsior must somehow refrain from giving a comic form to the worst excesses in the underlying data.

A sly wit is sometimes the saving grace of an ad hominem attack, but some topics are just too serious to ever be treated humorously. We scarcely want a creative system to make jokes about rape or the holocaust, but how and where do we draw the line? (Veale 2021) identifies two kinds of self-regulation for a creative system: *inner* and *outer* regulation. Any system using an inner regulator explores a modified search space that omits certain topics which can give rise to offense. So, by choosing not to put rape or the holocaust into its lexicon of possible hashtag referents, Excelsior becomes blind to those topics and will not use them in its comics. In this respect, the traditional knowledge bottleneck in symbolic systems can sometimes work to our advantage.

A system with an outer regulator does not explore a reduced search space, and so is capable, in principle, of treating sensitive topics in crass and insensitive ways. Instead, a filter is used to catch any potentially offenses before they can be shared with users or the public. For example, a "blocklist" might list the terms that a system must avoid. The filter is applied retroactively, so a system may explore, but not actually speak of, certain ideas. It follows that inner regulation makes more sense for a symbolic system whose knowledge is curated and pruned with care. Outer regulation, in contrast, is more suited to statistical systems that learn from real data. A hybrid system that uses a symbolic teacher to tune a statistical learner will use both kinds of regulation.

As such, Excelsior draws on both kinds of regulation to create comics that are informative, provocative and balanced. Yet, while the system is presently poised to fulfil its intended social function – automated intervention into ongoing debates – the foregoing ethical issues still give us sufficient pause to delay Excelsior's launch as an autonomous Twitter *bot*. An abundance of care is needed whenever one aims to balance potential harms against each other. Further testing is needed to quantify Excelsior's capacity to offend, since any system with the capacity to surprise may also shock and dismay.

Summary and Conclusions

The classic 1950s crime drama *Naked City* ended each episode with these words: "There are eight million stories in the naked city. This has been one of them." It seems natural to feel the same way about a large data set, such as a corpus of polarized views gathered from Twitter. This data does not tell a single story but many, and we must do justice to them all when we set out to creatively capture an overall sense of its contents.

Excelsior is a system for creating topical comics from an evolving social-media data set. It is a modular system that separates the planning of a comic – its plot, emotional cadence, and core opposition of views – from its visual rendering. For the former, Excelsior generates an XML specification of a comic which is human- and machine-readable, and for the latter it uses a bespoke renderer. The sequences of images in Figs. 1 and 2, for instance, have been generated by such a renderer. Its stories are composites, drawn from multiple sources and multiple related – or opposing – viewpoints. However, these composites still do justice to the data, by making explicit the narratives that connect different users, hashtags and tweets within and across echo chambers. Crucially, Excelsior balances the views in its comics, so that no single position is favored or goes unchallenged.

As a symbolic system, Excelsior relies on a number of explicit representations, which allow it to map hashtags onto topic-relative stances and emotions, and from there onto visual actions and textual dialogue. It is the logical coding of these representations that allows us to tightly maintain Excelsior's sense of balance. However, for the system to grow in expressive power, we need to make it learn for itself. So, mindful of the moral dilemmas that already attach to the symbolic model, and of how these might be exacerbated by inappropriate training data, we are tentatively exploring how the symbolic Excelsior might train its own statistical replacement.

Author Contributions

This paper is wholly the work of the principal author.

References

Alves, T.; McMichael, A.; Simões, A.; Vala, M.; Paiva, A.; and Aylett, R. 2007. Comics2d: Describing and creating comics from story-based applications with autonomous characters. In *Proceedings of CASA, the 20th Annual Conference on Computer Animation and Social Agents*.

Barberá, P.; Jost, J.; Nagler, J.; and Tucker, J. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science* 26(10):1531–1542.

Blaya, C. 2019. Cyberhate: A review and content analysis of intervention strategies. *Aggression and violent behavior* 45:163–172.

Eisner, W. 1985. *Comics & Sequential Art*. Tamarac, Florida: Poorhouse Press.

Frankfurt, H. 1986. On bullshit. *Raritan Quarterly Review* 6(2):81–100.

Gozalo-Brizuela, R., and Garrido-Merchan, E. 2023. Chatgpt is not all you need. a state of the art review of large generative ai models. *ArXiv* abs/2301.04655.

Hao, Y., and Veale, T. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Mind and Machines* 20:635–650.

Kurlander, D.; Skelly, T.; and Salesin, D. 1996. Comic chat. In *Proceedings of SIGGRAPH'96, the 23rd annual conference on Computer graphics and interactive techniques*. ACM. 225–236.

Loewenstein, J.; Raghunathan, R.; and Heath, C. 2011. The repetition-break plot structure makes effective television advertisements. *Journal of Marketing* 75(5):105–119.

Loewenstein, J. 2018. Surprise, recipes for surprise, and social influence. *Topics in Cognitive Science* 11.

McCloud, S. 1993. *Understanding Comics: The Invisible Art*. New York, NY: Harper Collins.

Melistas, T.; Siglidis, Y.; Kalogiannis, F.; and Manouach, I. 2021. A deep learning pipeline for the synthesis of graphic novels. In *Proceedings of ICCC-21, the 12th International Conference on Computational Creativity*. 256–265.

Proven-Bessel, B.; Zhao, Z.; and Chen, L. Y. 2021. Comicgan: Text-to-comic generative adversarial network. *ArXiv* abs/2109.09120.

Pérez y Pérez, R.; Morales, N.; and Rodríguez, L. 2012. Illustrating a computer generated narrative. In *Proceedings of ICCC-22, the 3rd International Conference on Computational Creativity*. 103–110.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, L. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21:1–67.

Rozin, P.; Rozin, A.; Appel, B.; and Wachtel, C. 2006. Documenting and explaining the common aab pattern in music and humor: Establishing and breaking expectations. *Emotion* 6:349–55.

Santana, B.; Campos, R.; Amorim, E.; Jorge, A.; Silvano, P.; and Nunes, S. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*.

Tang, L.; Cai, A.; Li, S.; and Wang, J. 2022. The naughtyformer: A transformer understands offensive humor. *arXiv* abs/2211.14369.

Veale, T., and Cook, M. 2017. *Twitterbots: Making machines that make meaning*. Cambridge, MA: MIT Press.

Veale, T. 2012. *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*. Bloomsbury Academic.

Veale, T. 2021. Walking the line in computational creativity: Paradox and pitfall in the mitigation of automated offense. In *Proc. of the 12th International Conference on Computational Creativity*. 132–139.

Veale, T. 2022. Two-fisted comics generation: Comics as a medium and as a representation for creative meanings. In *Proceedings of ICCC-22, the 13th International Conference on Computational Creativity*. 59–66.

Walsh, J. 2012. Comic book markup language: An introduction and rationale. *Digital humanities quarterly* 6(1).

Xie, Z.; Cohn, T.; and Lau, J. H. 2023. Can very large pretrained language models learn storytelling with a few examples? *ArXiv* abs/2301.09790.

On the power of special-purpose GPT models to create and evaluate new poetry in old styles

Piotr Sawicki[1], Marek Grześ[1], Fabricio Goes[3], Dan Brown[2], Max Peeperkorn[1], Aisha Khatun[2], Simona Paraskevopoulou[1]
[1] School of Computing, University of Kent, Canterbury, UK
[2] Cheriton School of Computer Science, University of Waterloo, Canada
[3] Computing and Mathematical Sciences Department, University of Leicester, UK
p.sawicki@kent.ac.uk, m.grzes@kent.ac.uk, fabricio.goes@leicester.ac.uk, dan.brown@uwaterloo.ca, m.peeperkorn@kent.ac.uk, aisha.khatun@uwaterloo.ca, simonaparaskevopoulou@gmail.com

Abstract

This study investigates the possibility of using GPT-3 models to generate high-quality poems in a specific author's style, through fine-tuning on datasets of poems accompanied by their metadata and automatically generated summaries. Our experiments show that a dataset of only 300 poems is sufficient to generate new poems in the style of a specific author. The evaluation was done through GPT-3 models fine-tuned for binary classification of GPT-3 outputs against the works of the original author. To establish the accuracy of GPT-3-based binary classifiers, we first tested them on a variety of texts and a range of classes, and found that their predictive accuracy is 99% on average. Using this method for poetry evaluation showed that the GPT-3 generated poems were indistinguishable from the original works of Walt Whitman and Rudyard Kipling in an average of 30% and 21% of the cases, respectively. This suggests that GPT-3 can be a useful tool in assisting authors, while further research is needed to turn it into an independent creator. Additionally, the workflow used in this study can be applied to other types of text and provides a way of using GPT-3 models for generating new content from user-provided summaries, when prompt engineering alone is insufficient.

Introduction

With the emergence of Large Language Models (LLMs), there has been tremendous growth, not only in Natural Language Processing (NLP) but also in Computational Creativity (CC). In particular, the GPT-series (Radford et al. 2018; 2019; Brown et al. 2020) is the main contributor to the progress. LLMs have an astonishing capacity for capturing and mimicking features from massive amounts of data. Although LLMs have attracted expected criticism (Birhane and Raji 2022; van Dis et al. 2023), e.g., with respect to stylistic reproduction (Floridi and Chiriatti 2020; Falk 2021), their reception has overall been positive (Brown and Jordanous 2022). Their remarkable generative capabilities warrant further exploration in Computational Creativity research (Dale 2021; Köbis and Mossink 2021).

Poetry creation, as a CC task, has been explored over the years using a wide variety of techniques (Lamb, Brown, and Clarke 2017; Oliveira 2017). There are expert systems (Misztal and Indurkhya 2014; Corneli et al. 2015), constraints-based approaches (Rashel and Manurung 2014; Toivanen et al. 2013), and linguistic models (Veale 2013; Hämäläinen 2018) that can imitate styles and produce novel poems. Moreover, machine learning techniques (Toivanen et al. 2014; Lamb and Brown 2019) and evolutionary approaches (Rahman and Manurung 2011) have achieved some success. Generating lyrics in the specific style with defined rhyme and meter constraints through Markov processes was explored in (Barbieri et al. 2012). Text-generation techniques have also been applied beyond poetry generation (Pachet and Roy 2014; Ens and Pasquier 2018) and can assist in the development of techniques to deliberately deviate from learned styles (Elgammal et al. 2017).

Large Language Models can generate high-quality texts, paragraphs, and short creative artifacts, such as poems or lyrics. The current applicability of LLMs goes beyond autonomous generation of novel artefacts, and practitioners use them in co-creative ways to explore, get inspired, or as a tool to overcome writer's block (Gwern Branwen 2019; 2022). Regardless, many creative tasks still require human moderation to filter out nonsensical responses and subpar results. To improve the quality of creative output of transformer-based systems, we need to explore what is possible, understand the challenges involved, and devise computer-based methods for verifying if the system is performing well. Likewise, it is crucial to determine differences in performance and associated costs between the various sizes and architectures of LLMs, allowing us to make informed decisions on model selection for the creative task at hand. In this paper, we present a preliminary exploration of these challenges, and offer current state-of-the-art recommendations.

As NLP research increasingly focuses on transformer-based approaches, computational creativity is starting to follow suit. Notable examples of using GPT-2 or BERT for poetry generation include fine-tuning GPT-2 for Chinese classical poetry (Liao et al. 2019), conducting an extensive human evaluation of GPT-2 generated English poetry (Köbis and Mossink 2021), experimenting with rigid constraints in poetry generation in both Chinese and English (Li et al. 2020), analysing the challenges of maintaining rigid stylistic constraints while using RNN and GPT-2 (Wöckener et al. 2021), exploring a transformative BERT-based approach to lyrics generation (Nikolov et al. 2020;

Oliveira 2021), and generating lyrics from GPT-2 and evaluating with BERT (Wesek 2019). Hämäläinen *et al.* (2022) experimented with combined encoder-decoder setup using RoBERTa and GPT-2 for modern French poetry generation. The methodology of human-computer co-creation of poetry have been explored in (Boggia et al. 2022), while (Stevenson et al. 2022) attempted to evaluate the creative abilities of GPT-3. Fine-tuning GPT-2 for poetry generation in the style of Emily Dickinson was explored in (Dai 2021). In (Lo, Ariss, and Kurz 2022) the authors have fine-tuned GPT-2 for limerick generation with special attention to maintaining the limerick rhyming scheme. (Chakrabarty, Padmakumar, and He 2022) worked on fine-tuning T0 and T5 LLMs for collaborative poetry generation.

In (Bons 2022) the author experimented first with generating song lyrics using prompt engineering with GPT-3, and subsequently with fine-tuning GPT-3 on a dataset of songs accompanied by songs' descriptions, artist biographies and song titles. The fine-tuning process allowed the author to generate higher quality lyrics than using prompt engineering only.

A similar approach from outside the field of computational creativity is the work of (Lee 2019; Lee and Hsiang 2020b; 2020a) who fine-tuned GPT-2 and BERT models for patent claim generation and evaluation. The authors fine-tuned GPT-2 on a dataset consisting of US patent claims, where each claim is accompanied by its summary and title. The system was subsequently able to generate patent claims from summaries provided by the user.

What those two works, song lyrics generation and patent claim generation, have in common is fine-tuning the models on the datasets where each entry is accompanied by its summary and other metadata. This allows the user to control the content of the output through a summary and other metadata provided in the prompt.

The latest version of GPT at the time of writing this paper, which is GPT-3.5 (`text-davinci-003`), is capable of generating poetry through prompt engineering alone. It can generate poems that are not only grammatically correct and have appropriate structure, but also tell a coherent story and can appear meaningful and evocative (Gwern Branwen 2022). However, the poems generated through prompt engineering alone, always appear to be written in the same style and use plain and simple language that lacks the unique personal perspective and emotional nuance that are hallmarks of human-generated poetry. Our initial experiments have shown that prompting GPT-3.5 to generate poems in the style of a specific author, e.g. Walt Whitman, does not lead to the desired outcome.

A well-structured poem is generated, and the narrative requested in the prompt is followed, but the style in an obvious way does not match the style of the requested author. One can assume that the works of all classical authors were part of the GPT-3.5 training dataset, but the style of a specific author cannot be reliably invoked through prompts. This issue is analyzed in detail in our companion paper (Sawicki et al. 2023).

Objectives and Methods

Our long-term objective is to build a system which can generate poems in the style of a specific author and with the subject and narrative provided by the user, thus allowing the user maximum control over the outcome. We fine-tune GPT-3 models on datasets of poems accompanied by their summaries and other metadata. We show that when GPT-3 is fine-tuned on the poetry of poet A, it will produce outputs in A's style even if the summary will request topics/content that the poet A has never written about. For example, we obtain poems written in the style of poet A about topics or content that appeared in the works of poet B.

Our second objective is to show that GPT-3 can also evaluate the correctness of style. We use GPT-3 to evaluate generated poetry using an automated approach motivated by the methodology presented in our previous work (Sawicki et al. 2022), where we have fine-tuned BERT models for binary classification of fragments from the works of an original author (Byron and Shelley in that case), against samples produced from GPT-2 models fine-tuned on the works of that author. The idea is that if the classifier cannot distinguish between those two categories, (i.e. the accuracy of the classifiers is around 50%), then the text has been successfully generated in the desired style.

This way of evaluation resembles the GAN argument: the produced item is regarded as "good" when the classifier cannot distinguish it from the set of items used to train the generator (Goodfellow et al. 2020). This approach, however, comes with a caveat: it can be argued that when the evaluation results are approaching 50%, instead of indicating the successful replication of the desired style, it may simply mean that the classifier is of poor quality. For that reason, we conduct a number of experiments to establish whether the fine-tuned GPT-3 models are reliable as text classifiers. We classify using fine-tuned GPT-3 models instead of BERT (which was the classifier used in our previous work (Sawicki et al. 2022)), because BERT requires large data sets to achieve good classification accuracy, and our poetry datasets are too small for that. We demonstrate that GPT-3-based binary classifiers achieve 99% accuracy when fine-tuned on only 200 samples per label.

The main contributions of this paper are:

1. We present a workflow that allows for generation of poems with a specific narrative and in a specific author's style through fine-tuning GPT-3 models. This approach could be extended beyond poetry to other categories of text, where prompt engineering alone does not give desired results.
2. We demonstrate that GPT-3 models fine-tuned for classification are highly accurate as text classifiers and can be used as a tool for poetry evaluation.
3. We provide a dataset of 2100 out-of-copyright poems (7 authors and 300 poems per author) where each poem is accompanied by a summary and a theme. This dataset can be used for further research on poetry generation.
4. We show new insights into the performance of various versions of GPT-3 models on poetry generation. The

smaller models (Ada and Babbage) produce results comparable to larger models (Curie and Davinci), thus considerably reducing the costs of fine-tuning GPT-3 for poetry generation and evaluation. This indicates that some tasks, like poetry generation, do not require the use of largest models.

The paper is organised as follows: Our dataset and the process of fine-tuning GPT-3 for poetry generation are presented in Part 1 on poetry generation. Poetry evaluation using GPT-3 as a classifier is the subject of Part 2 of the paper, where the results are also presented and analysed. The main findings of the paper are highlighted in Conclusion, where ideas for future work are also discussed.

Part 1—Poetry Generation

GPT-3 should not be thought of as a single system. It is available in four different sizes (Ada, Babbage, Curie and Davinci) and a multitude of fine-tuned versions. Fine-tuning is only available for the vanilla versions of the four sizes:

- Ada (2.7 Billion parameters),
- Babbage (6.7 Billion parameters),
- Curie (13 Billion parameters),
- Davinci (175 Billion parameters).

In this work, we use all four sizes of GPT-3 models fine-tuned separately for poetry generation and evaluation. We also use GPT-3.5 (`text-davinci-003`) to create summaries and themes of the existing poems.

While GPT-2 models can be fine-tuned on any text file, GPT-3 requires a fine-tuning dataset to be organized in a specific way, i.e., each entry must be in the form of:

```
{"prompt":"BODY_OF_PROMPT",
 "completion":"BODY_OF_COMPLETION"}
```

couplets. Therefore, GPT-3 cannot be fine-tuned on the dataset of poems alone. If the body of the poem is in the completion, we must decide what to put in the prompt.

While it is possible to fine-tune GPT-3 on a dataset where the prompt contains only the name of the author and the title of the poem, this does not give the user much influence over the narrative of the generated poem. As the body of the poem is the expected completion of the model, it is required that the instructions are provided to the model through the prompt by describing the narrative of the poem. Since this prompt is missing in the original dataset (and, in fact, in all publicly available datasets at the time of writing this paper), we use GPT-3.5 to create summaries for our corpus of poems, and then the original poems and their summaries are used to fine-tune instances of GPT-3 for poetry generation.

Data Preparation

To prepare our dataset, we scraped 2100 poems from publicly available sources (Project Gutenberg 2022; Poetry Foundation 2022). To lower the cost of running the experiments, we used only the poems that are more than 100 words and less than 500 words in length. This dataset contains the works of seven classical poets, and we have randomly selected 300 poems per author. These authors are:

- Ella Wheeler Wilcox (American, 1850–1919),
- Rudyard Kipling (English, 1865–1936),
- Emily Dickinson (American, 1830–1886),
- Lord Byron (English, 1788–1824),
- William Wordsworth (English, 1770–1850),
- Walt Whitman (American, 1819–1892),
- Thomas Hardy (English, 1840–1928).

We use only the works of authors who passed away more than 75 years ago due to copyright limitations. For all these poems, we generated summaries and main themes using GPT-3.5, and this process is explicated below.

Summary Generation For the generation of summaries and themes we used GPT-3.5 (`text-davinci-003`), which, at the time of writing this paper, is the most advanced GPT model dedicated to text generation.

Initially, each entry in the original dataset contains the following data: author, title, dates of author's birth and death (separated with a hyphen), author's country and finally, the body of the poem.

To generate the summary of the poem, we have used the following prompt:

```
"This is the poem:" +
BODY_OF_THE_POEM +
"This is the poem's summary:"
```

Theme Generation The rationale behind adding the main theme of the poem is to give an additional way of influencing the content of the generated poem. For example, we can provide a summary that describes a poem about love, and set the main theme as "Love". The same prompt could have the main theme set to "Sadness" thus affecting the poem's tone.

To generate the main theme of a poem (from the body of the poem), we have used the following prompt, which also includes the full list of themes that GPT-3.5 was selecting from.

```
"These are the categories: Mysticism,
Childhood, God, Love, Life, Art, Poetry,
Sadness, Despair, Depression, Death,
Religion, Nature, Beauty, Aging, Desire,
Travel, Dreams, Birth, War, Failure,
Immortality, Fantasy.
Choosing from these categories select
one that best describes this poem:" +
BODY_OF_THE_POEM
```

Poems Annotated with Summaries and Themes Each entry in our dataset is augmented with the main theme of the poem and the poem's summary. Thus, each entry in the final dataset has the following format:

```
<|startofauthor|>AUTHOR<|endofauthor|>
<|startofdates|>BORN - DIED<|endofdates|>
<|startofcountry|>COUNTRY<|endofcountry|>
<|startoftitle|>TITLE<|endoftitle|>
<|startofthemes|>THEME<|endofthemes|>
<|startofsummary|>
{BODY OF THE SUMMARY}
```

```
<|endofsummary|>
<|startofpoem|>
{BODY OF THE POEM}
<|endofpoem|>
```

The added tags are used to clearly delineate the specific items in each entry in the dataset. These tags are used both during fine-tuning of the GPT-3 models and during the generation of the poems later on. Our complete dataset that includes the original poems, their metadata, summaries, themes and tags is available on our GitHub repository[1].

Fine-tuning GPT-3 for Poetry Generation

OpenAI documentation (OpenAI-Documentation 2023) suggests using a dataset with a minimum of 500 entries (i.e. poems) for fine-tuning. Our dataset has only 300 entries for each specific author. This limitation is common in poetry analysis because, in general, poets do not produce a high volume of work. For this reason, we consider two approaches to fine-tuning GPT-3 on our data:

1. Fine-tune individual GPT-3 models for each author. Here, every model is based on 300 samples.
2. Fine-tune GPT-3 models on a combined dataset of all seven authors. Here, every GPT-3 model is fine-tuned on 2100 poems of 7 poets.

Additionally, we examine which GPT-3 model produces the best results when fine-tuned on our poetry dataset. The general guideline from OpenAI is to fine-tune smaller models for more epochs, and larger models for fewer epochs (given a dataset of a fixed size). We fine-tune Ada and Babbage models for four epochs, and Curie and Davinci for one epoch and four epochs when using 300 samples. When fine-tuning the models on 2100 samples, we fine-tune all models for four epochs.

The cost of fine-tuning GPT-3 for poetry generation at the time of writing this paper are as follows:

1. Davinci 300 samples 1 epoch - $6
2. Davinci 300 samples 4 epochs - $24
3. Davinci 2100 samples 4 epochs - $169

The cost of using Ada, Babbage, and Curie models are respectively 50, 40 and 10 times lower than using Davinci (OpenAI-Pricing 2023).

The summary of our fine-tuning configurations is presented in Tables 1 and 2. Table 1 shows that we fine-tune 6 models for every poet considered, and Table 2 shows that we create 4 models using the combined dataset of 2100 poems of all poets. All the hyperparameters of the GPT-3 models are left at their default values, and only the temperature was set to 1.

The following prompt-completion tuple structure is used for preparing the fine-tuning dataset for our GPT-3 models:

```
PROMPT:
<|startofauthor|>AUTHOR<|endofauthor|>
<|startofdates|>DATES<|endofdates|>
<|startofcountry|>COUNTRY<|endofcountry|>
<|startoftitle|>TITLE<|endoftitle|>
<|startofthemes|>THEME<|endofthemes|>
<|startofsummary|>
{BODY OF THE SUMMARY}
<|endofsummary|>
<|startofpoem|>

COMPLETION:
{BODY OF THE POEM}
<|endofpoem|>
```

[1]https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation

Model	Acronym	Fine-tuning epochs
GPT-3-Ada	4e	4
GPT-3-Babbage	4e	4
GPT-3-Curie	1e	1
GPT-3-Curie	4e	4
GPT-3-Davinci	1e	1
GPT-3-Davinci	4e	4

Table 1: Fine-tuning GPT-3 models for every poet separately. This method uses 300 samples per model.

Model	Acronym	Fine-tuning epochs
GPT-3-Ada	7A 4e	4
GPT-3-Babbage	7A 4e	4
GPT-3-Curie	7A 4e	4
GPT-3-Davinci	7A 4e	4

Table 2: Fine-tuning GPT-3 models for all poets. This method uses 2100 samples per model.

Generating Poems from the Fine-tuned GTP-3 Models

Because of the high cost of running GPT-3 on the OpenAI's servers (OpenAI-Pricing 2023), we limited our fine-tuning for poetry generation to two authors. We have randomly chosen Walt Whitman and Rudyard Kipling. This applies both to our single-author approach and when generating from the models fine-tuned on the seven authors' dataset. Given the information shown in Tables 1 and 2, and our fine-tuning on two poets, the number of fine-tuned models for poetry generation is 16 in our experiments (2 poets times 6 models in Table 1 plus 4 models in Table 2).

From each fine-tuned model, we generate 300 poems to be later used in evaluation in Part 2. In the case of models fine-tuned on the seven authors' dataset, we generate 300 poems in the styles of both of our selected authors. Generating a poem requires a summary and theme in the prompt. To make the poem generation exercise fair, we did not use summaries of the poems that were in any of the fine-tuning datasets. Instead, we summarised 150 poems for two additional authors, William Ernest Henley (English, 1849–1903) and Christina Rossetti (English, 1830–1894). We use those summaries as part of prompts for generating poems. Thus, for example, the prompt for generating poems in the style of Walt Whitman will have author, author's dates of life and author's country all set to Whitman's details, but the title,

AUTHOR: Walt Whitman
AUTHOR'S DATES: 1819-1892
COUNTRY: United States
TITLE: Paradise: In A Dream
THEME: Mysticism
SUMMARY: This poem describes a dream of Paradise, in which the narrator sees and hears beautiful things such as fragrant flowers, birds singing, a river of gold sand, the Tree of Life and the gateway to Heaven. The beautiful sight fills them with hope of seeing these heavenly things again one day.

Table 3: Example prompt used to generate the poem presented in Table 4. The complete tags are not shown here in order to save space, but are presented in the section on Poetry Generation.

theme, and summary will be selected randomly from either Henley or Rosetti.

We have found that apart from generating poems from our fine-tuned models using the prompt including all the metadata used above, it is also possible to add a starting fragment of the poem after the <|startofpoem|> tag. This allows the user to keep improving the poem in an iterative way. We do not report any results on this way of generating poetry, but we mention it here to emphasize the practical applicability of our methodology because it could form the basis of a practical iterative poetry generation system.

An example prompt used to generate poetry from our system is presented in Table 3, and the poem generated from this prompt is presented in Table 4. The poem is presented exactly as produced by our model, GPT-3 Curie fine-tuned on the works of Walt Whitman for 4 epochs. The provided summary is of the poem 'Paradise: In A Dream' by Christina Rosetti.

A DREAM of Paradise,
As in a chamber, in the middle of the woods, I stood alone;
A few faint stars shining in the window,
And a trace of red in the eastern window, giving out the light,
As if some fire had been quenched there.

As I so stood, suddenly appeared before me,
As it had been a huge transparent angel, in shape
not different from a man,
But of such exceeding beauty, that I knew him to be nothing less than a god.

He stood before me silent;
His eyes fixed on me, full, full of meaning looked the god;
Not to speak his name he said,
But beckoning with his left hand, he drew me after him into the woods,
Till we came to the edge of a river, shallow and rapid,
And stood by the banks of the river, where he stood silent.

Then he said–"Shall I show you Paradise?"
And he reached his right hand into the water,
And drew out with his fingers a handful of white sand,
And put the sand upon his left hand,
And made the sign of the cross upon me with the sand,
and gave me the handful of sand, and said,
"Look!"
And I looked–and behold a Paradise,
Birds, and fragrant flowers, and streams of gold sand,
And the River of Life, flowing and glittering,
With trees, and buildings, and places of refreshment,
And on the banks, beautiful women, and attractively dressed
men.

And this was the River of Life, which flowed and sparkled,
And the trees were decorated with light,
And the buildings with beautiful colors were suffused with light,
And the refreshment places were set forth with plate and glass,
And the home of glory and delight, and the place of perfect love.

And thence I saw all the earth,
And the cities, and the country with its farms,
And the roads and the teeming population,
And the farms with their fields and crops,
And the working of the machinery, and the noise of the whirr of the machines,
And the new products in the markets,
And the great ships at the wharves, and the steamers on the sea.

And I said, "How beautiful, how beautiful!"
And then I awoke, and I remembered the dream.

Table 4: Example poem generated by our system using the prompt shown in Table 3.

Part 2—Poetry Evaluation

Part 1 of this paper explains our method for poetry generation in the style of a specific poet. As explained above, we generated 300 poems using each of the 16 GPT-3 models that we fine-tuned. This part of the paper evaluates the quality of the generated poems.

We use automated evaluation following the methodology presented in our previous work (Sawicki et al. 2022), where BERT was used to distinguish between fragments of poems written by the original author and fragments of poems generated by fine-tuned GPT-2 models. Since BERT requires a large number of training data samples, in the above mentioned work, we classified 20-lines-long fragments of poems. In this paper, we evaluate entire poems, and this gives a limited amount of samples per label. As a result, using BERT-based classifiers is not viable. We show that GPT-3 can instead be used to classify poems when the amount of training data is small. We fine-tune GPT-3 to classify if a poem is written by the original author or generated by the fine-tuned GPT-3. This means that we have two orthogonal fine-tuning processes of the GPT-3 models in our paper.

Walt Whitman vs book on machine learning			
	Correct	Incorrect	Accuracy
Ada	199	1	99.5%
Babbage	**200**	**0**	**100%**
Curie	**200**	**0**	**100%**
Davinci	199	1	99.5%
Walt Whitman vs Leo Tolstoy			
	Correct	Incorrect	Accuracy
Ada	**200**	**0**	**100%**
Babbage	199	1	99.5%
Curie	**200**	**0**	**100%**
Davinci	196	4	98%
Leo Tolstoy vs book on machine learning			
	Correct	Incorrect	Accuracy
Ada	196	4	98%
Babbage	**200**	**0**	**100%**
Curie	189	11	94.5%
Davinci	180	20	90%
Walt Whitman vs Rudyard Kipling			
	Correct	Incorrect	Accuracy
Ada	196	4	98%
Babbage	**200**	**0**	**100%**
Curie	197	3	98.5%
Davinci	199	1	99.5%

Table 5: Results of evaluating the accuracy of GPT-3-based binary classifiers in Step 1.

Using GPT-3 for classification requires the implementation of the logit bias during inference. Logit bias is an optional parameter passed to GPT models during text generation. It modifies the likelihood of specified tokens appearing in the generated text. This parameter is represented as a mapping from tokens to their associated bias values, which are between -100 (a ban) to 100 (exclusive selection of the token). Moderate values between -100 and 100 will change the probability of a token being selected to a lesser degree. When this parameter is used, the bias changes the original probabilities of tokens generated by the model prior to sampling. Thus, passing the logit bias parameter for only two tokens, representing our classes “0” and “1”, both with a value of 100, will result in the models being able to output only these two tokens (OpenAI-Documentation 2023). Without this modification, the model may produce answers that will not indicate any of the classes, giving inconclusive classification results.

Our methodology for classification-based evaluation of poems consists of two steps:

1. Establishing the accuracy of GPT-3-based classifiers by conducting a series of experiments classifying various types of texts.
2. Evaluating GPT-3-generated poetry against the works of original authors using GPT-3-based classifiers.

Step 1—Establishing the Accuracy of GPT-3-based Classifiers

To establish the accuracy of the GPT-3-based classifiers, we trained classifiers on two-class text classification problems where the similarity between classes was ranging from completely dissimilar to increasingly similar. First, we classified Walt Whitman’s poetry against the extracts from a book on machine learning, ‘Reinforcement Learning, An Introduction’ by Sutton and Barto (2018). This was an example text that is very different from poetry. Then, we proceeded to classify Whitman’s poetry against fragments of prose from the Collected Works of Leo Tolstoy (Project Gutenberg 2022), and finally we classified Whitman’s poetry against the poetry of Rudyard Kipling as an example of two classes of text that are similar to each other. Additionally, we also classified extracts from the book on machine learning against fragments of prose by Tolstoy. Since all the poems in our dataset are between 100 and 500 words in length, when the samples from the book on machine learning or from the prose by Tolstoy are used, they have the random length between 100 and 500 words.

In all four of these experiments, the training/test split ratio is 2:1. The training dataset consists of 200 samples per label, and the test dataset consists of 100 samples per label. All the hyperparameters of the GPT-3 models used for classification are left at their default values, only the temperature was set to 0.

In order to determine which fine-tuned model produces the best results, for each experiment, we fine-tuned each of the four GPT-3 sizes: Ada, Babbage, Curie and Davinci. As per the instructions on the OpenAI website, we fine-tune Ada and Babbage classifiers for four epochs, and Curie and Davinci classifiers for one epoch.

The results of these experiments are presented in Table 5, and they show that there is almost no difference between the outcome from four different model sizes. This is a very useful finding, since it eliminates the need for using the largest Davinci-based models, thus greatly reducing experimental cost. Consistently, we find that GPT-3 can be a highly accurate text classifier. In almost every case, the accuracy of the classifiers was 98% or more, both on similar as well as dissimilar classes. The lowest score in all of these experiments was due to the Davinci model fined-tuned to classify the book on machine learning against the prose by Tolstoy, with the accuracy of 90.0%. The second worst performing model was Curie, also on the task of classifying the book on machine learning against the prose by Tolstoy, where it scored 94.5%. The scores for Ada- and Babbage-based classifiers were very similar. Overall, these experiments show that fine-tuned GPT-3 models are reliable as binary text classifiers to distinguish between different authors of poetry and different categories of text.

Since GPT-3-Babbage-based classifiers were most accurate on average, we chose the Babbage model as the basis for fine-tuning the classifiers for our poetry evaluation experiments below.

Step 2—Evaluating GPT-3-generated Poetry Against the Works of the Original Author Using GPT-3-based Classifiers

Now we describe our evaluation of GPT-3-generated poetry against the works of the original authors using GPT-3 as a classifier. We use the poems generated by our process of generating new poems described in Part 1 of the paper.

As in Step 1, the training/test split ratio for each classification was 2:1. Each training dataset consists of 200 samples per label, each validation dataset consists of 100 samples per label. Our evaluation defines a two-class classification problem, where label 0 represents generated poems, and label 1 denotes the works of the original author. The results are presented in Table 6. All the classifiers in this experiment are fine-tuned GPT-3 Babbage models, built as we explained above. Entries in the first column in the table tell us which fine-tuned GPT-3 model's output was label 0 (these are the poetry generator models obtained in Part 1), and this output was evaluated against the works of original author placed in label 1.

The results show that the accuracy of the classifiers varied from 61.5% to 87.5%. A higher accuracy indicates that the classifier was able to distinguish the GPT-3-generated poetry from the original works of the authors with a higher degree of success. On the other hand, a lower accuracy implies that the classifier struggled to distinguish between the two and that the GPT-3-generated poetry was similar to the original work of the human authors. An accuracy of 50% would mean that the classifier cannot differentiate between generated and original poems. The best result that we obtained on Whitman's style is 61.5%, and it demonstrates quite a high level of style preservation in the generated poems. The best result obtained on Kipling's style is 67%, which is less pronounced, but given the very high accuracy of this classification method in our calibration experiment reported in Table 5, one can argue that a large number of poems with well-preserved style was obtained on Kipling's style as well.

The results of classification show some differences in the level of style preservation between poetry generated from different models and different dataset sizes. Interestingly, we should note that poetry generated from Davinci-based models did not achieve the highest results for either of the authors. It means that the smaller GPT-3 models are sufficiently powerful to generate poetry in a selected style. We can speculate that the good performance of the smaller models may be due to the fact that the largest Davinci model may require more fine-tuning data to capture the style more faithfully.

The results in Table 6 also vary between the works of the two poets. Because of the high costs of running these experiments, we were limited to generating and classifying poetry of only two authors. Repeating these experiments with the works of other authors would provide more insights into style preservation of GPT-3 models, but our current results on the style of two poets indicate that our method has merit, and that it is possible to generate new poems in the style of a specific author.

In conclusion, the results of the experiments in Step 2 suggest that fine-tuning the smaller GPT-3 models is sufficient for the style preservation tasks, and it can be done effectively with a dataset of only 300 samples.

Our results show that there is no significant difference between models fine-tuned on 300 samples vs models fine-tuned on 2100 samples. However, fine-tuning on a dataset consisting of many poets' works could open the possibility of mixing poets' styles in the output. Instead of setting all the author's metadata in the prompt to, for example, Kipling's or Whitman's details, we could, for example, declare the author as "Rudyard Whitman". This approach, however, requires further research.

Walt Whitman GPT-3 vs Walt Whitman original			
Model	Correct	Incorrect	Accuracy
Ada 4e	127	73	63.5%
Ada 7A 4e	140	60	70%
Babbage 4e	131	69	65.5%
Babbage 7A 4e	134	66	67%
Curie 1e	150	50	75%
Curie 4e	**123**	**77**	**61.5%**
Curie 7A 4e	131	69	65.5%
Davinci 1e	144	56	72%
Davinci 4e	174	26	87%
Davinci 7A 4e	137	63	68.5%
Rudyard Kipling GPT-3 vs Rudyard Kipling original			
Model	Correct	Incorrect	Accuracy
Ada 4e	170	30	85%
Ada 7A 4e	147	53	73.5%
Babbage 4e	**134**	**66**	**67%**
Babbage 7A 4e	142	58	71%
Curie 1e	173	27	86.5%
Curie 4e	160	40	80%
Curie 7A 4e	150	50	75%
Davinci 1e	175	25	87.5%
Davinci 4e	161	39	80.5%
Davinci 7A 4e	163	37	81.5%

Table 6: Results of experiments in Step 2 where GPT-3-generated poetry is compared against the works of the original author. Entries in the first column in the table indicate which fine-tuned GPT-3 model's output was evaluated against the works of the original author. 7A refers to the dataset of seven authors (2100 samples), 1e or 4e indicate that the model was fine-tuned for one or four epochs, respectively.

These results should be interpreted with caution in the light of the fact that binary classifiers used are entirely black-box systems, i.e. we do not know how the classification was performed. However, having established the high accuracy of these classifiers in Step 1, we can, to some extent, rely on these results. Further investigation, especially including human evaluations, is necessary to thoroughly determine the quality of the GPT-3-generated poetry.

Discussion

Ventura (2016) suggests that to evaluate the generative system in the context of computational creativity, we should consider the factors of **novelty**, **value** and **intentionality**.

The system we proposed is capable of producing **novel** works, benefiting from the enormous amount of data contained in the original training dataset of the GPT-3 models. The prompting choices made by a human collaborator may also contribute to novelty.

As for **value**, the quality of the output was deemed indistinguishable from the works of the original authors on average in 25% of cases. Our workflow allows for some level of control over the output, and therefore can be a valuable tool for collaborative poetry creation.

Intentionality, however, stays entirely with the user: the fine-tuned GPT-3 poetry generator does not produce anything on its own, every generated poem is the result of user's input. The question of whether the computer can at all be deemed creative is a matter of an ongoing discussion (Guckelsberger, Salge, and Colton 2017), after all the machine will only do what it is told to do by its programmer and its user. Regardless of that, we can strive towards reducing the need for human input in producing the artifacts, or cherry-picking them from the multitude of system's outputs, and our system contributes toward these goals.

It is also worth considering the model as containing the intentionality of its creators, in building a general-purpose language system, amongst its implicit goals is the creation of high-quality topical poetry, since poetry is a major identifier of success for creativity in humans.

Our workflow of augmenting the dataset with summaries and themes, followed by fine-tuning GPT-3 models allows to generate poems in the specific author's **style**, which has proved impossible through prompt engineering alone.

The status of the overall task of style preservation as computational creativity task has been considered by Brown and Jordanous (2022), who give an overall positive answer. Certainly, building new poems in an existing style can delight readers, and in this sense alone, it surely provides novelty and value.

Conclusion

The main contributions of this paper are:

1. We create a dataset of out-of-copyright poems augmented with summaries and themes generated by GPT-3.5. This dataset can be used by researchers for further experiments with poetry generation.
2. We demonstrate that GPT-3 models fine-tuned on as few as 300 poems are effective poetry generators, able to generate poems in a desired style, and with a given theme and narrative. Smaller GPT-3 models fine-tuned for poetry generation perform as well as larger models (as evaluated by our method of binary classification), meaning that the task may not be as challenging as some other language tasks, and that it can be done fairly inexpensively.
3. We demonstrate that GPT-3 models fine-tuned for binary text classification on as little as 200 samples per label achieve on average 99% accuracy in separating those two classes, with smaller models performing equally good, or better, than much larger models.
4. Overall, we provide a system that is capable of generating poetry in user-controlled style and content. Our system can also be used in an iterative way: after providing the summary and metadata, the user can also provide a poem fragment, and continue generation from that point in the poem. Thus, our system can be a valuable "poet's assistant."

The workflow used in this paper might be a way to train specialised language models in general: to fine tune on an appropriate corpus where each item is accompanied by its summary, in order to generate new items from the user-provided summaries. We could see this workflow as a general-purpose way of taking advantage of the language fluency of GPT models, while also allowing for some focus on specific topics. This approach can still run afoul of standard concerns about artificial intelligence and knowledge, like Searle's Chinese Room argument (Searle 1980). More research is needed to explore this topic.

In future work, we will experiment with other ways of encoding the poems than by using summaries and themes.

We can also examine a poet's style change over the course of their career (Gervás 2011). Applying our current workflow to this task will require reducing the size of the fine-tuning dataset for poetry-generators even further, by splitting it into subsets, for example: 'EARLY WHITMAN', 'MIDDLE WHITMAN', 'LATE WHITMAN'. The question that will have to be answered first is: how small a dataset is sufficient to fine-tune GPT-3 for poetry generation?

In our dataset, the summary was almost always shorter than the poem. It would be interesting to test our approach on shorter poetic forms, like haiku, where the length of the summary would exceed that of the poem. It would be interesting to see how GPT expands the haiku into a summary, but also how it would generate the concise haiku from long summaries, especially to see if it can capture the structure of the haiku consistently.

Automated evaluation of poetry is an open problem. Our approach that uses GPT-3 is an encouraging one with a great potential for highly accurate results, but it is a "black-box" classifier. A promising alternative could be evaluation by virtual crowd, presented in (Goes et al. 2022), where the authors have examined the possibility of GPT-3 simulating the members of the jury that evaluates jokes, through answering the same questions that human evaluators were asked. The results were compared to the ground-truth of the human evaluation and found to be similar. This approach, however, has not been tested yet on poetry evaluation, and therefore, it is left for future research.

Author Contributions

Experimental design: PS with MG, FG, DB, AK, MP, SP; experimental implementation: PS with SP; writing: PS with MG, DB, FG, MP, AK, editing: MG, DB, FG, MP, AK.

Acknowledgements

We thank Anna Jordanous for useful discussions and feedback during the process of the development of this paper. We also thank the anonymous reviewers for their feedback and suggested references. The work of DB is supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada. MP is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds.

References

Barbieri, G.; Pachet, F.; Roy, P.; and Degli Esposti, M. 2012. Markov constraints for generating lyrics with style. In *ECAI*, volume 242, 115–120.

Birhane, A., and Raji, D. 2022. ChatGPT, Galactica, and the progress trap. *Wired*.

Boggia, M.; Ivanova, S.; Linkola, S.; Kantosalo, A.; and Toivonen, H. 2022. One line at a time—generation and internal evaluation of interactive poetry. In *Proceedings of the 13th International Conference on Computational Creativity*. The Association for Computational Creativity.

Bons, V. 2022. GPT-3 as a muse: generating lyrics. https://wingedsheep.com/songwriting-with-gpt-3/.

Brown, D. G., and Jordanous, A. 2022. Is style reproduction a computational creativity task? In *Proceedings of the 13th International Conference on Computational Creativity*. The Association for Computational Creativity.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, 1877–1901.

Chakrabarty, T.; Padmakumar, V.; and He, H. 2022. Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing. *arXiv preprint arXiv:2210.13669*.

Corneli, J.; Jordanous, A.; Shepperd, R.; Llano, M. T.; Misztal, J.; Colton, S.; and Guckelsberger, C. 2015. *Computational poetry workshop:* making sense of work in progress. In *Proceedings of the Sixth International Conference on Computational Creativity*, 268–275.

Dai, A. 2021. GPT-2 for Emily Dickinson poetry generation. Stanford University, Unpublished Dissertation.

Dale, R. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27(1):113–118.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: Creative adversarial networks, generating "Art" by learning about styles and deviating from style norms. In *Proceedings of the Eight International Conference on Computational Creativity*, 96–103.

Ens, J., and Pasquier, P. 2018. Caemsi: A cross-domain analytic evaluation methodology for style imitation. In *ICCC*, 64–71.

Falk, M. 2021. Artificial stupidity. *Interdisciplinary Science Reviews* 46(1-2):36–52.

Floridi, L., and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30(4):681–694.

Gervás, P. 2011. Dynamic inspiring sets for sustained novelty in poetry generation. In *ICCC*, 111–116.

Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd score: A method for the evaluation of jokes using large language model AI voters as judges. *arXiv preprint arXiv:2212.11214*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM* 63(11):139–144.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the" why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the Eight International Conference on Computational Creativity*, 128–135.

Gwern Branwen. 2019. GPT-2 neural network poetry. https://gwern.net/gpt-2.

Gwern Branwen. 2022. GPT-3 creative fiction. https://gwern.net/gpt-3.

Hämäläinen, M.; Alnajjar, K.; and Poibeau, T. 2022. Modern french poetry generation with RoBERTa and GPT-2. In *Proceedings of the 13th International Conference on Computational Creativity*. The Association for Computational Creativity.

Hämäläinen, M. 2018. Harnessing NLG to create Finnish poetry automatically. In *Proceedings of the ninth international conference on computational creativity*. Association for Computational Creativity (ACC).

Köbis, N., and Mossink, L. D. 2021. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in human behavior* 114:106553.

Lamb, C., and Brown, D. G. 2019. Twitsong 3.0: Towards semantic revisions in computational poetry. In *ICCC*, 212–219.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2017. A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts* 11(3):159–179.

Lee, J.-S., and Hsiang, J. 2020a. Patent claim generation by fine-tuning openai GPT-2. *World Patent Information* 62:101983.

Lee, J.-S., and Hsiang, J. 2020b. PatentTransformer-2: Controlling patent text generation by structural metadata. *arXiv preprint arXiv:2001.03708*.

Lee, J.-S. 2019. Personalized patent claim generation and measurement. *arXiv preprint arXiv:1912.03502*.

Li, P.; Zhang, H.; Liu, X.; and Shi, S. 2020. Songnet: Rigid formats controlled text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 742–751.

Liao, Y.; Wang, Y.; Liu, Q.; and Jiang, X. 2019. GPT-based generation for classical Chinese poetry. *arXiv preprint arXiv:1907.00151*.

Lo, K.-L.; Ariss, R.; and Kurz, P. 2022. GPoeT-2: A GPT-2 based poem generator. *arXiv preprint arXiv:2205.08847*.

Misztal, J., and Indurkhya, B. 2014. Poetry generation system with an emotional personality. In *ICCC*, 72–81.

Nikolov, N. I.; Malmi, E.; Northcutt, C. G.; and Parisi, L. 2020. Rapformer: Conditional rap lyrics generation with denoising autoencoders. In *Proceedings of the 13th International Conference on Natural Language Generation*, 360–373.

Oliveira, H. G. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th international conference on natural language generation*, 11–20.

Oliveira, H. G. 2021. Exploring a masked language model for creative text transformation. In *Proceedings of the Twelfth International Conference on Computational Creativity*, 62–71.

OpenAI-Documentation. 2023. OpenAI API documentation. https://platform.openai.com/docs/introduction.

OpenAI-Pricing. 2023. OpenAI pricing. https://openai.com/api/pricing/.

Pachet, F., and Roy, P. 2014. Non-conformant harmonization: the real book in the style of take 6. In *ICCC*, 100–107.

Poetry Foundation. 2022. https://www.poetryfoundation.org/.

Project Gutenberg. 2022. http://gutenberg.org/.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9.

Rahman, F., and Manurung, R. 2011. Multiobjective optimization for meaningful metrical poetry. In *Proceedings of the Second International Conference on Computational Creativity*, 4–9.

Rashel, F., and Manurung, R. 2014. Pemuisi: a constraint satisfaction-based generator of topical Indonesian poetry. In *ICCC*, 82–90.

Sawicki, P.; Grzes, M.; Jordanous, A.; Brown, D.; and Peeperkorn, M. 2022. Training GPT-2 to represent two romantic-era authors: challenges, evaluations and pitfalls. In *Proc. of ICCC*. Association for Computational Creativity (ACC).

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; and Khatun, A. 2023. Bits of grass: Does GPT already know how to write like Whitman?

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3):417–424.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*. The Association for Computational Creativity.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Toivanen, J.; Järvisalo, M.; Toivonen, H.; et al. 2013. Harnessing constraint programming for poetry composition. In *The Fourth International Conference on Computational Creativity*. The University of Sydney.

Toivanen, J.; Gross, O.; Toivonen, H.; et al. 2014. "The Officer Is Taller Than You, Who Race Yourself!": Using Document Specific Word Associations in Poetry Generation. In *Proceedings of the Fifth International Conference on Computational Creativity*. Jožef Stefan Institute.

van Dis, E. A.; Bollen, J.; Zuidema, W.; van Rooij, R.; and Bockting, C. L. 2023. ChatGPT: five priorities for research. *Nature* 614(7947):224–226.

Veale, T. 2013. Less rhyme, more reason: Knowledge-based poetry generation with feeling, insight and wit. In *ICCC*, 152–159.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24. Sony CSL, Paris.

Wesek, A. 2019. A comprehensive study of state-of-the-art word embedding algorithms for natural language generation. University of Kent, Unpublished MSc Thesis.

Wöckener, J.; Haider, T.; Miller, T.; Nguyen, T. T. L.; Pham, M. V.; Belouadi, J.; Eger, S.; et al. 2021. End-to-end style-conditioned poetry generation: What does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 57–66.

Monte Carlo Tree Search for Recipe Generation using GPT-2

Karan Taneja[1,2], **Richard Segal**[2], **Richard Goodwin**[2]
[1] School of Interactive Computing, Georgia Institute of Technology, GA, US
[2] Computational Creativity Group, IBM Research, NY, US
karantaneja@gatech.edu, rsegal@us.ibm.com, rgoodwin@us.ibm.com

Abstract

Automatic food recipe generation methods provide a creative tool for chefs to explore and to create new, and interesting culinary delights. Given the recent success of large language models (LLMs), they have the potential to create new recipes that can meet individual preferences, dietary constraints, and adapt to what is in your refrigerator. Existing research on using LLMs to generate recipes has shown that LLMs can be fine-tuned to generate realistic-sounding recipes. However, on close examination, these generated recipes often fail to meet basic requirements like including chicken as an ingredient in chicken dishes. In this paper, we propose RecipeMC, a text generation method using GPT-2 that relies on Monte Carlo Tree Search (MCTS). RecipeMC allows us to define reward functions to put soft constraints on text generation and thus improve the credibility of the generated recipes. Our results show that human evaluators prefer recipes generated with RecipeMC more often than recipes generated with other baseline methods when compared with real recipes.

Introduction

With the vast number of cooking recipes available online and the success of large language models (LLMs) such as GPT-2 (Radford et al. 2019), researchers have investigated the automatic generation of recipes by fine-tuning LLMs on large datasets of food recipes (Bié et al. 2020; Lee et al. 2020). Automatic food recipe generation can be used in the creative process of recipe design where a chef can explore ingredient combinations, take inspiration for new recipes, write recipe drafts, or learn about flavor patterns. LLMs trained to generate recipes can help chefs by generating multiple possible recipes, completing incomplete ingredient lists and recipe instructions.

Large language models such as GPT-2 can be fine-tuned with large recipe datasets such as Recipe1M+ (Marin et al. 2021) and RecipeNLG (Bié et al. 2020) to generate reasonable-looking recipes. However, the quality of the output recipes is often limited due to the presence of repetitive text and inconsistencies between different parts of the recipe. There are two main reasons for this: (i) LLMs generate text by producing one token at a time and appending it to the existing text, resulting in a high focus on local coherence but a lack of attention to the long-term vision of the given context, and (ii) fine-tuned models are typically trained on smaller datasets compared to the original model, which can lead to subpar generalization over the target domain.

In this paper, we propose a method to sample from fine-tuned LLMs using Monte Carlo Tree Search (MCTS) and simple reward functions that put soft constraints on text generation. These constraints aim to eliminate the irregularities in generated recipes, improving their plausibility and making them more appealing to human evaluators. Our method does not require any additional training after the domain-specific fine-tuning of LLMs and can easily be wrapped over an API that exposes the next token probabilities. Our work takes inspiration from Chaffin, Claveau, and Kijak (2022) where a discriminator network is used as the reward function with MCTS to generate text conditioned on specific classes.

Figure 1 shows a real example of (i) an ingredients list generated from a given recipe name and (ii) recipe instructions generated from a recipe name and ingredients list. In the first example, while the recipe name mentions "Shrimp," the baseline method fails to add shrimp to the ingredients list. In contrast, our proposed method, RecipeMC, includes shrimp in the ingredients list as expected. In the second example, the baseline method does not employ "lemon zest" in recipe instructions while RecipeMC uses all the given ingredients as anticipated. Further, RecipeMC also meticulously generates the complete names "pepper Jack cheese" and "green onion" unlike the baseline method which used the shortened versions "grated cheese" and "onion".

This paper makes three main contributions:

1. We introduce RecipeMC, a text-generation process based on MCTS that controls the output text with simple manually-defined reward functions to softly constrain recipes generated by a fine-tuned LLM model.
2. We provide evidence that MCTS outperforms traditional sampling methods for recipe generation using common automatic evaluation metrics.
3. We conduct human evaluations that show that our generated recipes are often indistinguishable from human-created recipes and humans may prefer recipes generated by RecipeMC.

Next, we review the related literature to place our work in the context of previous research in food analysis and recipe generation. We then describe our proposed recipe

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Recipe Name	John And Sarah's Best Sausage, **Shrimp**, Ham And Chicken Jambalaya
Ingredients (baseline)	4 celery ribs, chopped; 3-4 lbs chicken thighs; 1 teaspoon black pepper; 3 cups long grain white rice; 2 tablespoons garlic, chopped; 6 bay leaves; 1 teaspoon cayenne; 2 teaspoons salt; 1 lb smoked sausage; 1 teaspoon dried thyme; 1 cup green bell pepper, chopped; 1 cup ham, diced; 1 cup sweet onion, chopped; 2 cups chicken stock or 2 cups chicken broth; 4 cups water; 1 1/2 cups scallions, chopped; 8 tablespoons vegetable oil; 1 teaspoon oregano; 3 cups tomatoes, chopped (2 lb); 1 teaspoon white pepper
Ingredients (RecipeMC)	1 teaspoon dried oregano; 2 cups long grain white rice; 2 teaspoons cajun seasoning; 1 teaspoon dried basil; 1 teaspoon cayenne; 1/2 lb smoked ham, cut into 1/4 inch dice; 1 lb chicken, boneless, skinless, cut into 1/2 inch dice; 1 teaspoon dried thyme; 1 lb andouille sausage, cut into 1/4 inch slices; 1/2 lb **shrimp**, shelled, deveined and cut into 1/4 inch dice; 2 tablespoons olive oil; 2 teaspoons salt; 1 large onion, chopped; 4 cups chicken broth; 2 garlic cloves, minced

Recipe Name	Crab-Stuffed Corn Muffins Casserole
Ingredients	1 1/2 tablespoons mayonnaise; 1 tablespoon minced green onion; cooking spray; 1 cup cornmeal; 1/4 cup grated pepper Jack cheese; 1/2 teaspoon baking soda; 1/2 cup grated pepper Jack cheese; 1/4 teaspoon Worcestershire sauce; 1/2 cup melted butter; 1 cup buttermilk; 8 ounces cooked crabmeat; 2 eggs; salt to taste; 1 cup all-purpose flour; 1 teaspoon Asian chili paste (sambal); 1 teaspoon fresh grated **lemon zest**; 1/2 teaspoon salt
Instructions (baseline)	Preheat oven to 350 degrees F (175 degrees C). Spray a 12-cup muffin tin with cooking spray. Whisk together the eggs, buttermilk, and melted butter in a bowl. Stir in 1 cup of the flour, 1/2 of the cornmeal, the baking soda, 1/4 of the salt, the grated cheese, onion, mayonnaise, chili paste, and Worcestershire sauce. Mix in the crabmeat [...]
Instructions (RecipeMC)	Preheat the oven to 400 degrees F (200 degrees C). Spray 12 muffin cups with cooking spray. Combine cornmeal, flour, baking soda, and salt in a large bowl. Whisk together buttermilk, butter, eggs, pepper Jack cheese, mayonnaise, chili paste, green onion, **lemon zest**, and Worcestershire sauce in a separate bowl. Stir into the dry ingredients [...]

Figure 1: Examples of inconsistencies in generated recipes. In the first example, an important ingredient "Shrimp" is missing from the baseline ingredients list. In the second example, baseline instructions do not use the ingredient "lemon zest". Our method, RecipeMC, also refers to the complete names of the ingredients such as "pepper Jack cheese" and "green onion" rather than shortening them to "grated cheese" and "onion" respectively.

generation method RecipeMC including GPT-2 fine-tuning, MCTS, and the reward functions. This is followed by a section on experiments and results including the automatic evaluation of RecipeMC with three other baseline methods and a human evaluation study. Finally, we wrap up with a discussion, our ideas for future work, and concluding remarks.

Related Literature

Food and Recipe Analysis: Salvador et al. (2017) introduced the Recipe1M dataset containing over 1M recipes with 800K food images. With these recipes and corresponding food images, they trained text and image models to generate embeddings in a joint embedding space. These text and image models with their common embedding space were used to retrieve recipes from food images (im2recipe retrieval task) by matching the embedding of a given image with those of recipes in an existing database. Marin et al. (2021) extended this dataset to Recipe1M+ by further adding 13M food images. Min et al. (2017) proposed a Multi-Modal Multi-Task Deep Belief Network (M3TDBN) to learn from multi-modal content and multi-attribute information in the food domain for cuisine classification, recipe image retrieval, and ingredient and attribute inference from food images. Herranz, Min, and Jiang (2018) reviewed work involving multiple modalities in food analysis including text, images, location, and cuisine. Our paper focuses on coherent recipe text generation beginning with a recipe name to create ingredient lists and instructions.

Recipe Generation: Chef Watson (Varshney et al. 2019) was based on a Bayesian model over a knowledge-representation schema containing culinary information such as relations between ingredients, geolocations, and chemical composition of ingredients in terms of flavor compounds. Chef Watson could generate creative recipes whose quality was verified by professional chefs. Wang et al. (2020) proposed a method to generate recipes from food images using the Recipe1M dataset based on unsupervised extraction of paragraph structures and generating tree structures from images for a structure-aware generation. Bié et al. (2020) introduced the RecipeNLG dataset for recipe generation which was created by expanding the Recipe1M recipes with over one million new recipes. Lee et al. (2020) introduced RecipeGPT, a GPT-2 based recipe generation model and evaluation system with a user interface for examining the quality and encouraging experimentation. Reusch et al. (2021) introduced RecipeGM which generated recipes from a given list of ingredients (without quantities) using a hierarchical self-attention-based sequence-to-sequence model. This model was proposed by Fan, Lewis, and Dauphin (2018) for story generation where long-range dependencies are an important challenge. Overall, RecipeGPT performs consistently better than RecipeGM except at n-gram repetition, but note that RecipeGPT is a much bigger model as compared to RecipeGM. Antognini et al. (2022) developed a method to edit recipes through critiques where the latent representation of the recipe is modified using its gradient with respect to the critique. The critique is a list of desired ingredients in a recipe beginning with an initial recipe. In this paper, we propose RecipeMC which uses GPT-2 model for recipe generation with soft constraints on ingredient lists and instructions for improving the coherence and overall quality of the recipes. Our baseline methods use the same model as RecipeGPT, but use top-p sampling instead of top-k since it has been shown to consistently generate higher quality and more diverse text (Holtzman et al. 2020). We show that RecipeMC consistently performs better than these baselines on all metrics and human evaluation.

Constrained Text Generation: LLMs, such as GPT-2, struggle with maintaining the context of the prompt when generating structured responses. Researchers have explored constrained text generation using different methods. Zhang et al. (2020) and Hsieh, Lee, and Lim (2021) propose insertion-based transformer models to impose hard constraints that ensure the inclusion of given entities in the output text. To impose these constraints, these models progressively add tokens between the given entity tokens to generate text but also inadvertently constrain the order in which tokens are produced. Chaffin, Claveau, and Kijak (2022) propose a method for controlling generation using MCTS and a discriminator model as the reward function to generate text conditioned on a given discriminator class. This imposes a soft constraint on the output text, unlike insertion-based methods that enforce hard constraints. Our work takes inspiration from this paper but uses simpler manually-defined reward functions instead of a discriminator model. We linearly combine these reward functions to impose several soft constraints that encourage coherence and text quality in the recipes. This can be used in future research to allow human-in-the-loop collaborative recipe generation where humans prescribe the reward functions.

Recipe Generation

GPT-2 (Radford et al. 2019) is an LLM based on the transformer architecture (Vaswani et al. 2017) and can generate text conditioned on an initial prompt using next token prediction. GPT-2 was trained on a large corpus with millions of web pages. Fine-tuning a pre-trained LLM for specific tasks has been shown to have a significant advantage over training a model from scratch (Radford et al. 2018). For our method, RecipeMC, we fine-tuned a GPT-2 model using a large corpus of recipes collected from the internet. Each recipe contains three components: recipe name, ingredients list, and recipe instructions. Figure 2 shows the recipe format we use to fine-tune our language model. Here, `<|startofname|>` and `<|endofname|>` are special tokens used to denote the start and end of the recipe name respectively. Similar tags are also used for denoting the ingredient and instruction sections. We fine-tuned the GPT-2 model using the RecipeNLG dataset (Bié et al. 2020) which contains over 2.1 million recipes. We cleaned the dataset to remove text containing unwanted information such as text advertisements, empty or short ingredients, and instructions.

```
<|startofname|>Recipe Name<|endofname|>
<|startofingr|>
  Ingredient 1; Ingredient 2; ...;
  Ingredient n
<|endofingr|>
<|startofinst|>
  Instruction 1. Instruction 2. ...
  Instruction m.
<|endofinst|>
```

Figure 2: Recipe format used to fine-tune GPT-2. Special tokens define the start and end of each recipe section.

Prompt:

```
<|startofname|>
Chocalate Chip Cookies
<|endofname|>
```

Response:

```
<|startofingr|>
1/2 cup butter; 1/2 cup sugar;
1 large egg; 2 cups all-purpose flour;
1 cup semi-sweet chocolate chips;
<|endofingr|>
```

Figure 3: In the NAME→INGR task the system is given a recipe name and asked to generate a list of ingredients.

Similar to Lee et al. (2020), we also experimented with a multi-task learning setup where a model is trained to generate multiple possible orderings of name, ingredients, and instructions instead of only the default name-ingredient-instruction ordering. We did not see a decrease in perplexity over the test set, but rather a small increase, and decided to use the simpler single-order setup described in Figure 2.

For evaluation purposes, we split the recipe generation problem into two separate tasks. Figure 3 shows the NAME→INGR task where we prompt the system with the recipe name and ask the system to generate a list of ingredients. Figure 4 shows the NAME+INGR→INST task where we prompt the system with the recipe name and ingredients, and ask the system to generate the recipe instructions. We have split the task in this manner to simplify the evaluation of the two distinct types of structured text present in recipes. Note that running these two tasks one after another allows us to generate complete recipes from just the recipe name.

LLMs have a tendency to repeat text, especially the previous sentence at any point. These repetitions have a self-

Prompt:

```
<|startofname|>
Chocalate Chip Cookies
<|endofname|>
<|startofingr|>
1/2 cup butter; 1/2 cup sugar;
1 large egg; 2 cups all purpose-flour;
1 cup semi-sweet chocolate chips;
<|endofingr|>
```

Response:

```
<|startofinst|>
Preheat oven to 350F.  Combine all
ingredients in mixing bowl.  Mix in
chocolate chips. Place on baking sheet
and bake for 10 minutes.
<|endofinst|>
```

Figure 4: In the NAME+INGR→INST task the system is given a recipe name and a list of ingredients. The system is then asked to generate cooking instructions.

reinforcing effect — the probability of repeating a sentence successively increases with each repetition (Xu et al. 2022). Penalizing repetitions during inference can help mitigate this problem to some extent. Two common methods for inference time mitigation include (i) strictly disallowing n-gram repetitions and (ii) exponentially penalizing token repetitions (Shirish Keskar et al. 2019). Both these methods modify the token probabilities of LLMs during inference. A second major limitation of LLMs is their inability to reliably model long-range dependencies which leads to inconsistent text. For example, ingredients mentioned in the ingredients list may never be used in the generated recipe instructions (see Figure 1). Indiscriminately penalizing outputs for repetitions can further aggravate this problem because some repetitions, such as ingredients in recipes, naturally arise owing to their structure.

Monte Carlo Tree Search

MCTS is a search algorithm commonly used in AI agents for playing strategy games such as Chess, Go, and Checkers (Świechowski et al. 2022). It balances exploitation and exploration to efficiently search a large search space to find the path that maximizes a user-provided reward function. In text generation, MCTS can take a long-term view of the text generation process because it evaluates multiple possible paths emanating from the text prompt (Chaffin, Claveau, and Kijak 2022). The algorithm works by maintaining a partial search tree of all possible sentences as shown in Figure 5. The root node r of the tree represents the initial prompt $x_{1:r}$. The children of each node represent the possible continuations of the current sentence. MCTS starts with just the root node and then expands one node in each iteration of the algorithm. The figure shows the state of the algorithm after two iterations have been completed, and the root and “together” nodes have been expanded. The figure also demonstrates how the third iteration proceeds. In each iteration, MCTS

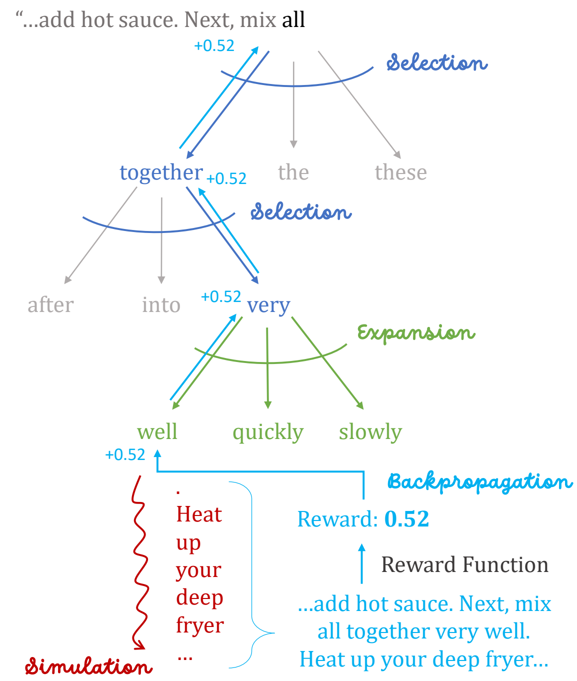


Figure 5: An illustrative example of RecipeMC in action for generating instructions. The initial prompt is “...add hot sauce. Next, mix all”. The four MCTS steps Selection, Expansion, Simulation and Backpropagation are shown and color coded.

performs the following main steps:

1. **Selection:** Starting from the root node, MCTS iteratively selects a child node to explore until it reaches a leaf node which will be expanded in the next step. The node selection is based on maximizing a variant of PUCB (Predictor + Upper Confidence Bound) (Rosin 2011; Silver et al. 2017) over the child nodes:

$$\mathrm{PUCB}(i) = Q(i) + c \cdot p(x_i|x_{1:i-1}) \frac{\sqrt{N}}{n_i + 1}$$

where the exploitation term $Q(i)$ is the average score from generating token x_i given $x_{1:i-1}$. $Q(i)$ is initialized with 0 whenever a node is created in the expansion step. It is updated for all selected nodes during backpropagation. The second term, c, is a constant that controls the weight of exploration and exploitation; n_i is the count of times that the child node i was visited; and $N = \sum_i n_i$ is the total number of iterations. The term $p(x_i|x_{1:i-1})$ is the predictor probability that serves as a prior. RecipeMC uses the output probabilities from the fine-tuned GPT-2 model for $p(\cdot)$. Note that PUCB(i) is also well-defined for unexplored child nodes that have $n_i = 0$. In Figure 5, the selection algorithm first picks “together” because it maximizes PUCB among the three options. It then selects “very” because it also maximizes PUCB. The selection step ends at “very” as it is a leaf node.

2. **Expansion:** From the selected leaf node l, we expand the tree by adding child nodes corresponding to the top-k tokens predicted by the fine-tuned GPT-2 model. We initialize the prior for each child by normalizing the probabilities over the top-k tokens to sum to one. In Figure 5, three nodes corresponding to the words "well", "quickly" and "slowly" are added. We used top-k here to limit the tree size as top-p will lead to an indeterministic size.
3. **Simulation:** We perform standard top-p sampling from the selected leaf node to generate the next t tokens giving us a text sequence $x_{1:l+t+1}$. In Figure 5, first, the leaf node "well" is selected using PUCB. Then, top-p sampling is used from "well" to generate the sequence "[period] Heat up your deep fryer [...]."
4. **Backpropagation:** We calculate the reward for text sequence $x_{1:l+t+1}$ and update aggregate scores $Q(\cdot)$ and increment $n_{(\cdot)}$ starting from node $l+1$ to the root node r. In Figure 5, the scores $Q(\cdot)$ are updated to accumulate a reward of 0.52, and $n_{(\cdot)}$ is incremented for the words "well", "very", "together", and the root node.

Finally, after repeating the above four steps Z times, one can decide the next token at the root node by choosing the child node with the highest $Q(r+1)$ or the node with the highest n_{r+1}. In our work, we select the node with the highest $Q(r+1)$ as the next token at the root. We used $Z = 20$, $c = 1$, $k = 50$, $p = 0.9$, and $t = 30$ without any further fine-tuning. To generate ingredients or instructions, this process is repeated until the corresponding end tag is reached.

Reward Functions

While previous MCTS work on LLMs has used discriminator models to guide the MCTS search (Chaffin, Claveau, and Kijak 2022), we use hand-designed soft constraints implemented by simple reward functions.

The reward functions discussed below use a predefined list of common ingredients such as milk, eggs, butter, chicken, etc. which we call *constituents*. To create the constituents list, we applied the NYT Ingredient Phrase tagger[1] (NYT-IPT) over the ingredient phrases in the RecipeNLG dataset. The NYT-IPT allows tagging of quantities, units, ingredient names, and comments within ingredient phrases. We extracted the set of ingredient names from all ingredient phrases in the ingredients lists and filtered them to remove any constituents with non-alphabet characters or stop words ("and", "or", etc.). Further, we filtered out constituents that could be decomposed into other constituents present in the list. The final list has 2,122 constituents.

Since NAME→INGR and NAME+INGR→INST tasks have very different outputs, they require different reward functions. The reward functions discussed below were designed to address several structural shortcomings we found in recipes that were generated without MCTS. For instance, several generated recipes without MCTS failed to include the key ingredient that defines the recipe such as not including chicken as an ingredient in "Chicken Masala." In RecipeMC, we used the following three reward functions for NAME→INGR task:

- **Name & Ingredients Coherence:** This function rewards the model for using ingredients names present in the recipe name. For example, for the recipe *Chocolate Apple Pie*, the function rewards outputs with ingredients *Chocolate* and *Apple*. We first search for constituents in the recipe name. If $z > 0$ constituents are found, we search these z constituents in the generated ingredients list and, say, z_f are found. The reward value is $z_f/z \in [0,1]$ if $z > 0$, and 1 otherwise.
- **Constituents Repetition Penalty:** This function penalizes any repetition of constituents in the ingredient list. Let p be the sum of the number of times a constituent is repeated. Note that we do not count the first occurrences. Similarly, let q be the sum of the number of times the ingredient phrases separated by ";" are repeated. Then, the reward is given by $e^{-p-q} \in (0,1]$.
- **Closing Ingredients List:** This function rewards the <|endofingr|> token generation. If the tag is found, the reward is 1, and 0 otherwise.

The total reward q is defined as the weighted sum of the reward functions ($q = \sum_i w_i r_i$) where weight $w_i \in (0,1)$ is assigned such that $\sum_i w_i = 1$. This ensures $q \in [0,1]$ since each function value $r_i \in [0,1]$. We used the weights 0.30, 0.45, and 0.25 for the above three functions respectively without any hyper-parameter fine-tuning.

Similarly, we also use three reward functions for the NAME+INGR→INST task:

- **Ingredients & Instructions Coherence:** Similar to *Name & Ingredients Coherence*, this function rewards the use of constituent names present in the ingredients list when generating recipe instructions. We first search for constituents in the ingredients list. If $z > 0$ constituents are found, we search these z constituents in the generated instructions (say z_f are found). The reward value is $z_f/z \in [0,1]$ if $z > 0$, and 1 otherwise.
- **Special Characters Repetition Penalty**: We observed the model tends to repeat some characters like "!" and "-" because they may be repeated in some training recipe instructions. If the sum of occurrences of these characters is s, the reward is given by $e^{-s/S} \in (0,1]$. We used $S = 3$ to avoid excessive penalization for using these characters.
- **Closing Recipe Instructions:** This functions rewards the <|endofinst|> tag. If the tag is found, the reward is 1, and 0 otherwise.

For generating instructions, we used the weights 0.50, 0.20 and 0.30 for the above three functions respectively without any hyper-parameter fine-tuning.

Experiments and Results

We create a LLM for our experiments by fine-tuning GPT-2 on our cleaned RecipeNLG data. We use the same LLM for all our experiments. We separately evaluate RecipeMC on the NAME→INGR task and the NAME+INGR→INST task.

[1] https://github.com/nytimes/ingredient-phrase-tagger

Table 1: Automatic evaluation results for NAME→INGR task using different sampling methods. The down arrow (↓) indicates that lower is better.

Sampling Method	Coherence	F_1-Score	Perplexity↓	ROUGE-1	ROUGE-2	BLEU	Repetition↓
Ground Truth	0.451	-	2.934	-	-	-	0.667
Top-p	0.443	0.572	4.173	0.457	0.200	0.155	1.724
+ No 4-gram Repetition	0.444	0.562	5.150	0.456	0.198	0.144	1.641
+ Repetition Penalty	0.413	0.548	6.754	0.407	0.135	0.115	0.711
RecipeMC	**0.513**	**0.597**	**3.961**	**0.505**	**0.242**	**0.210**	**0.192**

We compare RecipeMC with following three baseline sampling methods commonly used with LLMs:

- **Top-p Sampling:** Top-p sampling, also known as nucleus sampling (Holtzman et al. 2020), uses tokens with the highest probabilities that cumulatively add up to the nucleus size p and zeroes out the probability of other tokens. This is an adaptive version of top-k sampling where exactly k tokens with the highest probabilities are considered independent of their cumulative probability. This method has been shown to generate more diverse and interesting text than other greedy approaches.
- **Top-p Sampling with Repetition Penalty:** To prevent the repetition of tokens, the output logit values of repeated tokens are divided by a parameter $\theta > 1$ and the distribution is re-normalized(Shirish Keskar et al. 2019). We use the recommended value $\theta = 1.2$ along with top-p sampling as a baseline.
- **Top-p Sampling with No n-gram Repetitions**: This baseline method also uses top-p sampling but forbids repetition of n-grams. To ensure that no n-gram is repeated, we can search for the last n-1 generated tokens in the sequence generated so far and find the list of tokens that follow them. These tokens should not be generated to ensure that there are no n-gram repetitions. This method enforces a strict constraint unlike the repetition penalty but it allows repetitions of n-1-grams or smaller sequences without any penalization. We used $n = 4$ in our experiments to allow for some margin in repetition.

Automatic Evaluation

We compare RecipeMC with the three baselines on several standard metrics. We used 1,000 test recipes with ground-truth ingredient lists and instructions and compare the generated ingredient lists and instructions for each method with the ground truth. For some metrics where ground truth is not required, we also report the values for ground truth as an oracle reference. For the NAME→INGR task, we gave the recipe name from each test recipe as a prompt and sampled the LLM's output ingredients list using all three baseline methods and RecipeMC till the model generated the end tag. Similarly, for the NAME+INGR→INST task, we gave the recipe name and ingredients from each test recipe as a prompt and sampled the output instructions. The results from the automatic evaluation for NAME→INGR and NAME+INGR→INST tasks are summarized in Table 1 and 2 respectively.

Coherence: We compare the *coherence* of different methods by comparing constituents present in the generated ingredients list and the recipe name, and between instructions and the ingredients list. This definition is inspired by the definition of coherence given by Lee et al. (2020). For NAME→INGR task, we define coherence with *Name & Ingredients Coherence* function defined earlier. Similarly, for NAME+INGR→INST task, we define coherence with the *Ingredients and Instructions Coherence* function. The results in Tables 1 and 2 show that RecipeMC achieves the highest coherence, surpassing even the ground-truth recipes. These results confirm that the Coherence reward functions have the desired effect of improving the coherence of the recipes generated with MCTS.

F_1-Score: (only for NAME→INGR task) In order to compare the quality of generated ingredients with respect to ground truth, we search for constituents in the ground-truth ingredients list and calculate the average precision and recall for the generated ingredients list. The F_1-Score is the harmonic mean of average precision and recall. The results in Table 1 show that RecipeMC has the highest F_1-score among all methods confirming that the *Name & Ingredients Coherence* reward also leads to higher ingredients accuracy with respect to the ground truth.

Perplexity: The perplexity of a sample, defined with respect to a language model, measures the surprise of the model in seeing the given example. It is defined as the exponent of cross-entropy over the sequence of tokens in a given text. Even though LLMs are trained to minimize the cross-entropy loss, the text-generation process or the inference method can influence the perplexity of a sampled text. For this metric, we only consider the perplexity of the generated part of the output and mask the output probability for the prompt text. The result in Tables 1 and 2 show that RecipeMC generates recipes with the lowest perplexity, but still more than that of the test dataset. This improvement over other baseline methods is because MCTS allows us to look ahead before the next token generation and to avoid tokens that later lead to poor outputs with lower reward values. It is interesting to note that our method did not explicitly reward lower perplexity, but its low perplexity is a consequence of the soft constraints imposed by the reward functions and the additional search performed by MCTS.

ROUGE & BLEU: ROUGE or Recall-Oriented Understudy for Gisting Evaluation (Lin 2004) is a set of metrics to evaluate the quality of text in summarization and machine-translation tasks. It measures the quality of output text by

Table 2: Automatic evaluation results for NAME+INGR→INST task using different sampling methods. The down arrow (↓) indicates that lower is better.

Sampling Method	Coherence	Perplexity↓	ROUGE-1	ROUGE-2	BLEU
Ground Truth	0.486	4.115	-	-	-
Top-p	0.709	7.948	0.338	0.102	0.067
+ No 4-gram Repetition	0.690	8.441	0.339	0.103	0.069
+ Repetition Penalty	0.416	11.680	0.301	0.072	0.044
RecipeMC	**0.768**	**7.337**	**0.362**	**0.115**	**0.080**

measuring the overlap, i.e. recall, precision, and accuracy of n-grams between the output text and reference texts. We use ROUGE-1 and ROUGE-2 F_1 values to measure unigram and bigram overlap between the generated recipe texts and the original recipe texts. BLEU or Bilingual Evaluation Understudy (Papineni et al. 2002) metric was proposed to measure the quality of machine-translation systems and has been shown to have a high correlation with human judgment. It combines the precision of n-grams where $n = 1, 2, 3, 4$, and a brevity penalty for generating output text shorter than reference text. We measure the BLEU score of output ingredients and instructions by comparing them to the original ones in the test set. The results in Tables 1 and 2 show that RecipeMC generates recipes closest to the ground-truth recipes. Note that the same GPT-2 model was used for each method, but the sampling process used by RecipeMC led to higher-quality ingredient lists and instructions.

Repetition: (only for NAME→INGR task) Repetition is defined as the average number of repetitions of constituents in the ingredients list. Zero value indicates that no constituents were repeated in the ingredients list. Table 1 shows that RecipeMC leads to minimum repetitions, but the repetition value for RecipeMC is even lower than that of the ground-truth recipes. This may indicate that we are over-penalizing the constituent reuse in the ingredients list.

Output Length: The average character length of output ingredient lists and instructions are reported in Table 3. We observe that ground-truth recipes have the shortest length. RecipeMC has the shortest length among all sampling methods since we reward the generation of the end tag. It is interesting to note that top-p sampling with repetition penalty has the shortest length among the baseline methods for ingredient lists but a much higher length than the other baselines for instructions. We observed on inspection that the repetition penalty led to a quicker generation of the end tag for ingredients but created a *blabbering* effect for instructions where the model generates extremely elaborate instructions and occasional unrelated text about alternate or complementary recipes using some ingredients that were not present in the given ingredients list.

Overall, RecipeMC outperformed the baseline methods while balancing several objectives which led to good-quality recipes. Using top-p sampling without any constraints outperformed other baseline methods except for its highly repetitive text as constituents were repeated 1.7 times on average as compared to 0.7 times for ground truth. The top-p sampling baselines with No 4-gram Repetition and Repetition Penalty led to lower repetitions as compared to top-p sampling, but it reduced coherence, F_1-Score, ROUGE, and BLEU values and increased perplexity. RecipeMC achieved the best of both worlds with higher F_1-score, ROUGE, and BLEU values when compared to ground-truth recipes with the least repetitive text. The smaller average length of output for RecipeMC (Table 3) also confirmed that it was able to succinctly capture more relevant information. As discussed next, our human evaluations also confirmed that humans prefer recipes generated by RecipeMC.

Table 3: Average character length of the ingredients list and instructions for different sampling methods.

Method	Ingredients	Instructions
Ground Truth	167	240
Top-p	247	485
+ No 4-gram Repetition	248	484
+ Repetition Penalty	233	545
RecipeMC	190	441

Human Evaluation

To evaluate each method on NAME→INGR task using human evaluation, we created a *Recipe Turing Test* where a human evaluator looks at two possible ingredient lists, the real one and the generated one, for a given recipe name, and their task is to **choose the generated ingredients list** among these. For a fair comparison, we uniformly shuffled the real and generated recipes to show them on the left or right side. Also, evaluators were not aware that four different methods were being evaluated. We perform a similar test for NAME+INGR→INST task where human evaluators are asked to **choose the generated instructions** based on a recipe name and the associated ingredients list.

We randomly sampled 50 recipes for each of the four methods and created a total of 200 binary-choice questions. The evaluators see 10 randomly-chosen questions (without replacement) for NAME→INGR task and 5 questions for NAME+INGR→INST task on a separate test. Note that all evaluators did not take both the tests. The results of the human evaluation for the two tasks are shown in Tables 4 and 5. *Real* and *Gen.* columns count the number of times the real and generated recipes were selected by the evaluator. P(Incorrect) is the probability that humans incorrectly identified the real recipe as the generated one. This is calculated

Table 4: Human evaluation results on the NAME→INGR task. Data was collected from 147 evaluators who answered 10 questions each. Evaluators were asked to identify the generated ingredients (**Gen.**), hence P(**Incorrect**) = P(**Real**) = **#Real / (#Real + #Gen.**).

Method	**Real**	**Gen.**	P(**Incorrect**)
Top-p	175	185	0.4861
+ No 4-gram Repetition	179	200	0.4723
+ Repetition Penalty	183	180	0.5041
RecipeMC	201	167	**0.5462**
Overall	738	732	0.5020

Table 5: Human results on the NAME+INGR→INST task. Data was collected from 83 evaluators who answered 5 questions each. Evaluators were asked to identify the generated ingredients (**Gen.**), hence P(**Incorrect**) = P(**Real**) = **#Real / (#Real + #Gen.**).

Method	**Real**	**Gen.**	P(**Incorrect**)
Top-p	51	42	0.5484
+ No 4-gram Repetition	67	62	0.5194
+ Repetition Penalty	36	65	0.3564
RecipeMC	57	35	**0.6196**
Overall	211	204	0.5084

as follows:

$$P(\text{Incorrect}) = \frac{\#\text{Real}}{\#\text{Real} + \#\text{Generated}}.$$

The results in Tables 4 and 5 show that human evaluators are most likely to believe that RecipeMC ingredient lists and instructions are human-generated when compared to those produced by the other baselines. Human evaluators believed recipes generated by RecipeMC to be human-generated more often than the next best method, top-p sampling with Repetition Penalty, for NAME→INGR task with a p-value of 0.128. Similarly, RecipeMC outperforms top-p sampling on the NAME+INGR→INST task with $p = 0.164$. While we did not observe statistically significant ($p < 0.05$) improvements with RecipeMC results, we observe that results from human evaluation correlate well with results from automatic evaluation presented in Tables 1 and 2.

We also observed that top-p sampling with repetition penalty performed at par with other baselines for NAME→INGR task but performed worse than other baselines on the NAME+INGR→INST task with $p = 0.007$. This can be attributed to its unusually long instructions as discussed earlier and shown in Table 3.

Juxtaposing these results with average lengths shown in Table 3, we observe that human evaluators prefer recipes with shorter length among the generative methods, but do not prefer the shortest recipes, i.e the real recipes, over RecipeMC for both ingredients and instructions. This confirms that humans did not heavily rely on text length. Assuming that human evaluators select outputs at random when recipes are equally good, it is surprising to see that the ingredient lists generated by RecipeMC are perceived to be more human-like than the original (human-written) recipes with $p = 0.042$ and instructions are perceived to be more human-like than original recipes with $p = 0.014$. We believe that this is because (i) RecipeMC repeats ingredients far less than the original recipes, as shown in Table 1, and (ii) it uses complete ingredient names consistently, as shown in Figure 1, and indicated by coherence values higher than ground truth shown in Tables 1 and 2.

Discussion and Future Work

MCTS and manually defined reward functions provide an effective way to control the ingredients and instruction generated by LLMs without requiring additional training. This approach offers flexibility and can enable users to generate recipes that adhere to specific constraints. It can also be valuable for interactive recipe editing. Users can present partial recipes to the system and ask the system to fill in the blanks. Personal constraints such as being sugar-free, low sodium, or vegetarian can be added to the reward function. Users can iteratively prompt the system with different combinations of ingredients and collaboratively create new recipes. We plan to explore interactive recipe generation in future work to measure novelty and creativity through user studies with amateur and professional chefs.

The ability of MCTS to look ahead during the generation of each token coupled with heuristic-based reward functions interestingly led to lower perplexity and higher similarity to ground truth recipes without directly optimizing for these properties. This finding is interesting from the perspective of text generation and suggests that using MCTS with LLMs may be applicable to a wider set of applications beyond structured text generation.

Conclusions

We presented a new sampling method, RecipeMC, that combines LLMs, MCTS, and custom reward functions to generate recipes that are often indistinguishable from human recipes for the same dish. We have shown that our method outperforms common text-generation approaches for LLMs for this task on a variety of automatic generation metrics. We conducted a *Recipe Turing Test* and found that users preferred RecipeMC ingredients about 55% of the time and RecipeMC generated instructions 62% of the time as compared to human-generated ingredients and instructions. These evaluations show that MCTS combined with manually-defined reward functions can be an effective tool for recipe generation with LLMs such as GPT-2.

Author Contributions

Author 1 was in charge of writing the manuscript, planning the study, building the system, and setting up the programs for automatic and human evaluation. Author 2 ran the experiments and compiled and analyzed the results. All three authors contributed to algorithmic design, writing the manuscript, planning the study, and conducting the human evaluation.

References

Antognini, D.; Li, S.; Ai, M.; Faltings, B.; and Mcauley, J. 2022. Assistive Recipe Editing through Critiquing. *arXiv:2205.02454*.

Bié, M.; Gilski, M.; Maciejewska, M.; Taisner, W.; Wi´sniewski, D. W.; and Ławrynowicz, A. 2020. RecipeNLG: A Cooking Recipes Dataset for Semi-Structured Text Generation. In *Proceedings of the 13th International Conference on Natural Language Generation*, 22–28.

Chaffin, A.; Claveau, V.; and Kijak, E. 2022. PPL-MCTS: Constrained Textual Generation Through Discriminator-Guided MCTS Decoding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2953 - 2967*, 2953–2967.

Fan, A.; Lewis, M.; and Dauphin, Y. 2018. Hierarchical Neural Story Generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, 889–898. Association for Computational Linguistics (ACL).

Herranz, L.; Min, W.; and Jiang, S. 2018. Food recognition and recipe analysis: integrating visual content, context and external knowledge. *arXiv:1801.07239*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; Choi, Y.; and Allen, P. G. 2020. The Curious Case of Neural Text DeGeneration. In *Proceedings of the International Conference on Learning Representations*.

Hsieh, L.-H.; Lee, Y.-Y.; and Lim, E.-P. 2021. ENCONTER: Entity Constrained Progressive Sequence Generation via Insertion-based Transformer. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 3590–3599.

Lee, H.; Shu, K.; Achananuparp, P.; Prasetyo, P. K.; Liu, Y.; Lim, E. P.; and Varshney, L. R. 2020. RecipeGPT: Generative Pre-training Based Cooking Recipe Generation and Evaluation System. In *WWW-20 - Companion Proceedings of the World Wide Web Conference*, 181–184. Association for Computing Machinery.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Association for Computational Linguistics*, 74–81.

Marin, J.; Biswas, A.; Ofli, F.; Hynes, N.; Salvador, A.; Aytar, Y.; Weber, I.; and Torralba, A. 2021. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43(1):187–203.

Min, W.; Jiang, S.; Sang, J.; Wang, H.; Liu, X.; and Herranz, L. 2017. Being a supercook: Joint food attributes and multi-modal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia* 19(5):1100–1113.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI*.

Reusch, A.; Weber, A.; Thiele, M.; and Lehner, W. 2021. RecipeGM: A Hierarchical Recipe Generation Model. In *Proceedings of the IEEE 37th International Conference on Data Engineering Workshops*, 24–29. Institute of Electrical and Electronics Engineers Inc.

Rosin, C. D. 2011. Multi-armed Bandits with Episode Context. *Annals of Mathematics and Artificial Intelligence* 61:203–230.

Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Ofli, F.; Weber, I.; and Torralba, A. 2017. Learning Cross-modal Embeddings for Cooking Recipes and Food Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shirish Keskar, N.; Mccann, B.; Varshney, L. R.; Xiong, C.; Socher, R.; and Research, S. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *arXiv:1909.05858*.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; Van Den Driessche, G.; Graepel, T.; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676):354–359.

Świechowski, M.; Godlewski, K.; Sawicki, B.; and Mańdziuk, J. 2022. Monte Carlo Tree Search: a review of recent modifications and applications. *Artificial Intelligence Review* 56(3):2497–2562.

Varshney, L. R.; Pinel, F.; Varshney, K. R.; Bhattacharjya, D.; Schörgendorfer, A.; and Chee, Y.-M. 2019. A Big Data Approach to Computational Creativity. *IBM Journal of Research and Development* 63(1):1–7.

Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Proceedings of the Conference on Neural Information Processing Systems*.

Wang, H.; Lin, G.; Hoi, S. C.; and Miao, C. 2020. Structure-Aware Generation Network for Recipe Generation from Images. In *Proceedings of the European Conference on Computer Vision*, volume 12372 LNCS, 359–374. Springer Science and Business Media Deutschland GmbH.

Xu, J.; Liu, X.; Yan, J.; Cai, D.; Li, H.; and Li, J. 2022. Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation. In *Proceedings of the Conference on Neural Information Processing Systems*.

Zhang, Y.; Wang, G.; Li, C.; Gan, Z.; Brockett, C.; and Dolan, B. 2020. POINTER: Constrained Progressive Text Generation via Insertion-based Generative Pre-training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 8649–8670. Association for Computational Linguistics (ACL).

Creative Data Generation: A Review Focusing on Text and Poetry

Mohamad Elzohbi, Richard Zhao
Department of Computer Science, University of Calgary
Calgary, Alberta, Canada T2N 1N4
{melzohbi, richard.zhao1}@ucalgary.ca

Abstract

The rapid advancement in machine learning has led to a surge in automatic data generation, making it increasingly challenging to differentiate between naturally or human-generated data and machine-generated data. Despite these advancements, the generation of creative data remains a challenge. This paper aims to investigate and comprehend the essence of creativity, both in general and within the context of natural language generation. We review various approaches to creative writing devices and tasks, with a specific focus on the generation of poetry. We aim to shed light on the challenges and opportunities in the field of creative data generation.

Introduction

Data refers to information that can be stored and processed by a computer. It can take many forms and can be generated both naturally, such as through the thoughts and ideas in a person's mind, and artificially, such as through the use of machine-based models. In particular, computer scientists use the term data to refer to anything that can be stored in the computer's memory called binary data. The process of generating data can be (1) fully automated, like using a generative model to generate poetry from topic words, (2) semi-automated, where the output is a collaboration between the machine and the human, such as a poem draft generated with human refinements (Lamb, Brown, and Clarke 2017), or (3) entirely manual such as using a text editor to write a poem. The concept of creativity has historically been difficult to define and has not been considered seriously in AI, as it was thought that machines need first to be capable of possessing thoughts and experiencing emotions (Colton, Wiggins, and others 2012). With the advent of deep learning and transformer models, natural language generation (NLG) techniques have become more advanced and the possibility of generating creative output seems more viable.

This paper focuses on the intersection between creative data generation and NLG with an emphasis on poetry generation. First, we address the difficulties in defining creativity and present our own perspective on the key elements of creativity. Second, we review relevant metrics used and how they are relevant to the proposed criteria, then we evaluate NLG models based on those criteria. Third, we provide an overview of the practical creative applications of text generation tasks, reviewing some of the most recent work in this area. We focus on the poetry generation task, examining the methods and models employed.

Creativity Paradox

Creativity is a complex and often-debated concept that is far from being well-defined. While many tend to focus on the creativity of the end result, the process by which it is created may also play a role in determining its level of creativity. When the steps involved in generating an output are clearly defined and easy to replicate, the output may be seen as less creative, even if it was initially considered as such. For example, the creativity of Johannes Vermeer, one of the most well-known Dutch artists in western history, was put onto the table of controversial discussions when Philip Steadman suggested that Vermeer was using optics in the painting process (Steadman 2002; Teller 2014). Furthermore, the English painter Hockney (2006) argued that most painters since the 15th century secretly used their knowledge of optical science in the painting process. It could be questioned as to why certain painters would choose to keep their techniques secret if this does not raise doubts about the authenticity of their artistic talent and creativity. One possible explanation is that artists desire for their work to be perceived as unique and distinguished from craft-based forms of expression. They may view their work as constituting a higher form of artistic expression, and therefore seek recognition as elite artists rather than as craftsmen, who are considered to have a lower social status (Markowitz 1994). This indicates that the level of creativity in the output can be significantly impacted by the degree of unpredictability. Hence, can photography be considered creative? Paul Delaroche, a classical painter, believed that the emergence of photography marked the death of painting. However, painters disagree and see painting as expressive and original, as it is not simply a mirror reflection of the real world. (Crimp 1981). On the other side, pictorialists argue that they can express their vision through photography by adding their own touches to a real photograph (Hertzmann 2018). Consequently, how valid is it to attribute creativity to machines if they are simply following a predetermined set of rules written by humans? In fact, even when we have a clear understanding of the design of a machine, we may not fully comprehend the internal processes that occur during its training. Does the hint of anonymity allow for the possibility of attributing creativity to them? Does revealing the process diminish the perceived originality of its output? Do we, as humans, understand creativity based on beautifulness or wonderfulness of the output?

Shneiderman (2007) summarizes the creative process into three main schools: (1) structuralists: who outline the creative process in four stages: (i) gathering information (preparation), (ii) forming new connections (incubation), (iii) finding sudden insights (illumination) and (iv) verifying and refining those insights (verification). (2) inspirationalists: who believe that creative insights can be achieved through sketching, visualization and meditation. (3) situationalists: who see creativity as a social process. In Boden (2004)'s point of view, creativity must induce (i) new, (ii) surprising, and (iii) valuable ideas or artifacts. As a situationalist, Csikszentmihalyi (1997) asserts that novelty is not enough to be considered creative, but the work must be accepted by the relevant field. Hertzmann (2018) argues that machines are merely tools in the hands of artists. These tools are not always predictable but not inherently creative. Hertzmann argues that just as we do not attribute creativity to a brush when watercolors flow on a canvas or to nature and animals when they form beautiful patterns and stunning structures, although animals have brains just like humans, we should not attribute it to machines.

Ultimately, the concept of creativity appears to be complex and multifaceted, making it seem impossible to develop a cohesive paradigm. However, we can see recurring themes and intersecting dimensions that we will summarize in the following: (see Figure 1)

1. **Originality**: To be considered creative, an output must strike a balance in originality. To achieve this, a creative model must avoid lacking originality (plagiarism) especially when learning from creative works. The greater the originality in the output, the more creative it is considered to be. However, too much originality results in confusion, disconnection, and excessive refinement (preciosity).

2. **Unpredictability**: An output that is obvious and predictable is generally seen as banal and boring. The greater the surprise factor in the output, the more creative it is considered to be. However, the output that is completely unpredictable may be perceived as nonsensical and lose its creativity. To achieve a balance between surprise and coherence, a creative model may need to learn a human-like probability distribution and optimize its decoding strategies for creative expression.

3. **Sociability**: We define sociability as the density of meaning and imagery in the output in relation to its appropriate domain. Poets and painters, for example, use language and colors, respectively, to compact complex ideas and emotions into a limited space, such as a verse or a canvas. A work that is semantically condensed provides more material for the artist or critic to engage with and assess. If the work is less sociable, it may be seen as redundant. However, it is important for the output to strike a balance, as an output that is overly dense may be complex and difficult to comprehend.

Creative NLG: Aspects

Discussions about NLG in the past have centered around determining what constitutes as *real* NLG (Reiter 2016). As the field progresses, the focus has shifted towards determining what constitutes as *creative* NLG. We believe that a creative NLG model must exhibit a sufficient level of originality, sociability, and unpredictability.

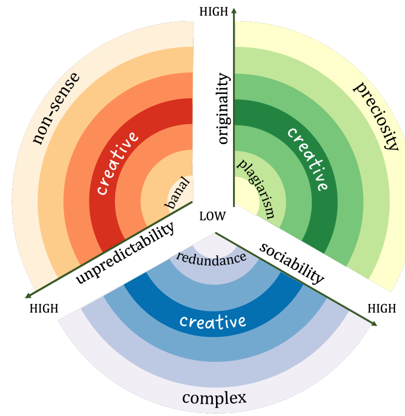


Figure 1: Creativity Dimensions.

Despite the significant advances in NLG, the study of the originality in language models has received relatively little attention. Originality in language can take various forms, such as phrasal, thematic, semantic, or stylistic originality, among others. Although probabilistic models may lead to the generation of original phrases, there is also the risk of over-fitting and copying from the training corpus. Plagiarism in creative language is particularly noticeable, as the output deviates from the normal use of language. This has been, recently, a concern in other creative fields, such as in music generation (Yin et al. 2021). Brooks and Youssef (2021) have noted that there is currently no standard automatic test for originality in NLG and proposed an approach to test the original use of language to identify plagiarism violations using a phrasal counting approach in the ground truth. However, this approach is limited to measuring the phrasal originality which is also important to identify phrases that diverted from their connotation and lost their imagery due to repetitive use in ordinary speech, such as dead metaphors.

Shannon (1951) conducted experiments to predict language redundancy by asking humans to guess the next character in an English passage. Shannon assumed that humans sample from a probability distribution based on their previously learned letter-sequence probabilities and introduced the concept of language entropy.

$$H = -\sum_i p\left(w_i\right)\, \log\left(p\left(w_i\right)\right)$$

where $p(w_i)$ is the probability for the ith n-gram to occur. Paisley (1966) studied the effect of several factors, including structure, on the redundancy of the language and found that prose is more redundant than poetic language. A more redundant language is less dense hence less sociable. Due to form restrictions in poems, it seems at first glance that the population of the allowed text must decrease from that of prose, but Manin (2021) postulates that poetic devices expand the population due to the relaxation of language norms. A well-known metric that is used to test NLG output quality is perplexity, which is directly related to entropy. A study on the correspondence between perplexity and human evaluation revealed that creative output requires perplexity that is not too high and not too low (Keukeleire 2020).

A recent study (Berns and Colton 2020) has introduced the concept of active divergence, which refers to the ability of a model to actively diverge from the training data to achieve a more original and diverse output, as opposed to the traditional method of training the model to perfectly mimic the training data. However, it may be more challenging for a language model to diverge from the training data while still adhering to the rules of grammar, semantics, and syntax. The sampling methods also play an essential role in the output predictability. Applying random sampling leads to a random non-sensical output, and greedy decoding tends to prefer banal or repetitive output. Holtzman et al. (2020) noted that beam search decoding in NLG models may result in less surprising and more monotonic output compared to the high variability in human choice of words. In order to increase unpredictability while keeping randomness low, they proposed Nucleus Sampling, which randomly sample from the top-k most probable words and set a probability threshold p to ensure that sampling does not occur from peaked or flat distributions. Only the top-p tokens are considered where $V^{(p)}$ is the smallest set of tokens such that:

$$\sum\nolimits_{x \in V^{(p)}} P\left(x|x_{1:i-1}\right) \geq p$$

The level of creativity in language models is influenced not only by the inference algorithms, but also by how similar the machine language modeling is to human language modeling. Deep learning models, particularly those that use transfer learning, have achieved impressive results on NLG tasks. Auto-regressive (AR) models, such as Generative Pre-Training models (GPT/2/3), generate text by sampling the next token from a probability distribution. Although humans may use causal language when they speak, they often try to modify their speech when they attempt to write creative text. Masked Language Models (MLM), such as BERT, predict masked tokens in a sentence based on context, which is similar to how humans try to guess a better word after removing it from a sentence and considering the context. However, these models will not automatically identify the words to remove and replace. Guided generation allows the model to be responsive and adaptable, and to accept or reject criteria, which may be a step towards a more social model. A study on the steerability of Generative Adversarial Networks (GANs) (Jahanian, Chai, and Isola 2019) shows that they can extrapolate yet are limited to the diversity of the data set. Previous NLG models have been limited by fixed or monotonically increasing sequence lengths, but Levenshtein Transformer (Gu, Wang, and Zhao 2019) and Edit-based Transformer (Xu and Carpuat 2021) allow for dynamic length changes. However, current large language models (LLMs) tend to overlook the integration of the unique elements associated with various forms of creative writing, like poetic diction, in their training process.

Creative NLG: Devices and Tasks

As discussed earlier, creativity in natural language appears in the original, rich and unusual use of language. Figurative language utilizes figures of speech to modify the literal meanings of words and produce imagery, or utilizes figures of sound to produce an appealing form. These rhetoric devices are used in creative writing in special ways to ensure novelty, richness and unpredictability. This includes similes, metaphors, hyperboles, sarcasm, humor, rhythms and rhymes, among others. These techniques are combined to shape the elements of various types of creative writing such as poetry and prose novels. While most research has focused on the detection of these literary devices, it was until recently that researchers showed interest in models that generate such devices. The employment of MLM and AR based models, along with the versatile BART model that amalgamates the strengths of both approaches and the use of COMET (Bosselut et al. 2019), designed for generating commonsense knowledge, underpinned the observed trend. Next, we will review recent work and its impact on downstream NLG tasks before we introduce poetry generation.

Metaphor: Automated metaphor generation has only recently received increased attention (Tong, Shutova, and Lewis 2021). MERMAID (Chakrabarty et al. 2021) finetunes BART with an automatically generated parallel corpus using an MLM model. The MLM will replace metaphorical verbs with their literal complement using the symbol relation from COMET. Then, they use a generator-discriminator approach to transform literal expressions to metaphorical ones with a verb-replacement objective and a top-k sampling strategy. The generated sentences are scored using a RoBERTa-based metaphor detection model to favor sentences with higher quality metaphorical verbs. Stowe et al. (2021) propose a similar model that replaces literal verbs with metaphorical ones guided by conceptual mapping. The authors propose two methods, one by training a frame embedding model and the other by finetuning BART using a generated parallel corpus similar to the one used in MERMAID but tagging each sentence with FrameNet frame labels. A limitation of these models is that they only generate metaphorical verbs and assume a given context.

Simile: Chakrabarty, Muresan, and Peng (2020) finetune BART on an automatically built corpus that transforms similes to their literal compliments using the "has property" relationship provided by COMET to generate similes from a given literal sentence by replacing a word by a novel simile. A limitation of the proposed model is that it explicitly replaces only an adjective or an adverb from a given input. Zhang et al. (2021a) propose a simile insertion approach based on the general context of the sentence without replacing words. The approach consists of two main stages: (1) detecting an appropriate simile insertion position in the input sentence using a BERT-based model (2) generating a simile at that position using a standard transformer decoder. Both models are trained jointly in a multi-task learning setup.

Hyperbole: MOVER (Zhang and Wan 2022) finetunes BART on a dataset of hyperboles retrieved from an online corpus using a BERT-based hyperbole detection model. The model is trained on a mask-filling task where potential hyperbolic tokens are identified, masked, and then regenerated based on their part-of-speech (POS) n-gram and ranked by an unexpectedness score. To generate a hyperbole, the model masks a non-hyperbole sentence span using a POS n-gram and then selects the highest-ranked of multiple possible hyperbole candidates using a BERT-based hyperbole ranker. HypoGen (Tian, krishna Sridhar, and Peng 2021) focuses on clause-level hyperboles with a specific pattern of the form (A1 is so A2 that B is D). To train this model, the authors collected and annotated a dataset of hyperboles from Reddit posts (HYPO-Red), selecting a subset that contains the target pattern and analyzing the relationships between clauses. They then used COMET and reverse-COMET models to generate candidate clauses with commonsense and counterfactual relations, respectively, and used BERT-based and neural-based classifiers to select the top-k candidates.

Sarcasm: To generate sarcasm, Chakrabarty et al. (2020) negate or replace evaluative words with their antonyms and provide a context to emphasize the semantic incongruity between the intended sarcasm and the context. Several context candidates are retrieved using a concept derived from the COMET model and an online corpus. The candidates are ranked by a RoBERTa model for semantic incongruity and the highest scored are concatenated to the sarcastic sentence. Oprea, Wilson, and Magdy (2021) argue that it is not sufficient to only negate the literal meaning of a sentence to produce sarcasm and propose a framework that is based on the implicit display theory of sarcasm to generate a sarcastic response to an input. The framework identifies an ironic environment by negating the event in the input utterance. A COMET-based rule-based method is used to produce an insincere negative attitude response based on the input event and one of the following relations: the action needed to perform the event, the attribute needed to perform the action, the user reaction, and the effect of the action on the user.

Puns: Yu, Tan, and Wan (2018) use a conditional LSTM model to generate homographic puns given two senses of a pun word. They use a joint beam search decoding with a forward and backward generation centered by the target pun word. Both sequences are then concatenated to form the final pun sentence. AMBIPUN (Mittal, Tian, and Peng 2022) generates homographic pun sentences using GPT-3 to generate related context words given two different senses of a pun word. The context words are then combined using a T5 model to generate candidate sentences and a BERT-based model to rank and pick a final pun sentence.

Multi-Figurative: Creative writing tasks need to employ various literary techniques in the generation. It is more practical to have one multi-figurative generation model to do so. Lai and Nissim (2022) propose a multi-figurative model that can transform between literal or figurative forms to another figurative form. The authors first finetune a BART model on a denoising objective to infuse multi-figurative sense into the model, then they fine-tune the model on a literal to figurative paraphrasing objective with a parallel data corpus and another time with a figurative to figurative paraphrasing objective with a cross attention layer to leak the target figure of speech to the encoder. For inference, they either transform the input text to the target figure of speech directly or to a literal form first and then to the target figurative form.

Impact on NLG Downstream Tasks

Generating creative language plays an essential role in many text-to-text NLG tasks, especially in creative writing tasks such as poetry and prose novels or stories. It also improves the quality of downstream tasks such as machine translation, dialogue generation, text summarization, style transfer, and even other data-to-text applications.

Storytelling is to generate an open-ended text that conveys to the reader a comprehensive story. It is one of the creative writing tasks written in prose form as opposed to poetry. Storytelling needs to account for soft and hard constraints such as the topics, plots and characters among other aspects to keep the story consistent and coherent. AI-aided approaches were used (Goldfarb-Tarrant, Feng, and Peng 2019) to collaboratively write and plan story plots. Plug-and-Play Language Model (PPLM) (Dathathri et al. 2019) is used for controllable story generation by plugging an attribute classifier on top of a GPT-2 transformer that checks how similar the next token is to a given topic. The classifier will be exposed to the history of latent representations of the generated words, then repeatedly perform backward and forward passes through the classifier and the generator calculating gradients each time, updating the representations, and handing them back to the generator. This will steer the generation process towards the topics at inference time with minimal training. The work was limited to assistive stories generation, completing a story skeleton given a story theme or sentiment. Brahman and Chaturvedi (2020) introduce multiple emotion-aware GPT-2-based models coupled with the COMET model to generate stories given a title and an emotional arc of the protagonist. Their work was the first to generate stories with an emotional trajectory plan.

Text Style Transfer is gaining an increasing popularity in natural language generation (Garbacea and Mei 2020). Variational Auto-Encoders (VAEs) were used to rewrite modern text in Shakespeare style (Mueller, Gifford, and Jaakkola 2017). Riedl (2020) used a transformer-based model (XL-Net) to generate a parody of lyrics, changing the lyrics while preserving the rhythm and syllable count.

Machine Translation Low (2011) examines the challenges and strategies involved in translating jokes and puns. The author argues that while some forms of humor are easily translatable, others, such as those that rely on wordplay and cultural references, can be particularly difficult. To effectively translate this kind of humor, the paper suggests using a variety of strategies, such as sense transferring, semantic leaps and cultural substitution.

Dialogue Generation can be modulated to generate creative dialogue content such as improv games, battle rapping, interactional jokes and conversational narratives, among others. In fact, recent chatbots such as LaMDA (Thoppilan et al. 2022) and ChatGPT[1] have shown the ability to write

[1] openai.com/blog/chatgpt/

and complete jokes, poems, and stories in conversations.

Text Summarization is the task of condensing a longer text preserving the most important information within its content. Similarly, creative language generation can be seen as summarizing broad ideas and beautiful imagery using words. In fact, iPoet (Yan et al. 2013) formulates the poetry generation task as a text summarization task. Given users' intents, write a poem that obeys poem requirements by retrieving terms out of a poetry corpora.

Data-to-text generation: Creative data-to-text tasks are analogous to how humans describe, in creative writing, what they see or hear from natural beauty, scenery, natural soundscape, etc. Loller-Andersen and Gambäck (2018) used Inception, ConceptNet and LSTM models to generate image-inspired poems. Liu et al. (2018) proposed a multi-adversarial CNN-RNN-based GAN model to caption images with a poem. Uehara et al. (2022) proposed transformer-based encoder and decoder models to generate emotional narratives from visual embedding extracted from images using CNNs. Achlioptas et al. (2021) demonstrated the power of utilizing their explained emotion-captioned image data set in enhancing language models to express and explain emotions triggered by artistic images. In addition, Chen and Lerch (2020) proposed a SeqGAN-based model that generates a matching line of lyrics with an input melody.

Poetry Generation

Modeling poetry is more difficult than language modeling as it requires the machine not only to understand but also to use language as a creative tool. As previously discussed, creativity is not yet well-defined, and it is uncertain if computers can achieve it.

Definition Before discussing how to generate poetry, we first need to define poetry, which is equally difficult as defining creativity itself. In fact, there are as many ways to describe poetry as there are people (Murfin and Ray 2009). Poetry is often defined as a type of writing that uses distinctive style, rhythm, and language to convey intense feelings and ideas [2] or the writing that concentrates imaginative experience to elicit a specific emotional response through meaning, rhythm, and sound patterns [3]. According to Milic (1970), poetry is the writing that violates the logical sequence and semantic categories of prose. Both prose and poetry may use literary devices such as metaphors, similes, ironies, and puns to express ideas and evoke emotions. Poetry, however, employs specific, but not essential, devices such as rhyme, rhythm, and meter. Meanwhile, the most notable devices found in a poem are its versification and form coherence.

Creativity Dimensions Various definitions of poetry highlight the three dimensions of creativity previously mentioned, delivered through language. A poem is a language artifact that requires the use of language to express meaning, and thus must adhere to linguistic rules to some extent while also allowing for violations, commonly referred to as poetic license. Originality and unpredictability can be observed through the surprising deviation from prose and normal speech rules, as well as through the use of poetic devices and euphony. The social aspect of poetry can be emphasized through the concentration of ideas, imagery and emotions within the boundaries of verses and stanzas, as well as the ability for the poem to be subject to criticism.

Forms English poetry can be classified based on various poetic characteristics. Notably, rhythm and rhyme schemes are often used to classify poems. The most popular forms include: (1) *Free Verse*: which has no constraints on a specific form. (2) *Haiku*: a short form of poetry that follows a 5-7-5 syllabic pattern. (3) *Sonnet*: characterized by specific rhyme and meter schemes. Shakespearean sonnet is considered one of the most well-known with 3 quatrains and a couplet: ABAB CDCD EFEF GG rhyme scheme. (4) *Blank Verse*: consists of unrhymed but metered verses. (5) *Limerick*: consists of five lines with seven to ten syllables with a verbal rhythm and the first, second, and fifth lines rhyming.

Goals The most common goals of writing poetry according to Preminger (2016) are to imitate reality, attain special effects on the readers, communicate emotions, and be art for art's sake. Similarly, the automatic poetry generation must represent adequate computational goals. Milic (1970) upheld in an early view that the possible usefulness of computer poetry is to influence the doer (designer/end-user) to learn more about language, poets and poetry.

Poetry Generation Techniques

Automatic poetry generation was documented as early as 1959 with a word salad approach by manually enhancing permutations of poem words (Lutz 1959). Since the few similar early attempts, the automatic poetry generation topic was inactive until the 1990s, where some works started getting attention in the field (Oliveira 2009). Early poetry generation techniques can be categorized (Gervás 2002) into: template-based, generate-and-test, evolutionary, and case-based reasoning. Manurung (2003) groups poetry generation approaches into: word salad, template/grammar-based, form-aware and poetry generation systems. Manurung argues that the first three methods do not satisfy the definition of poetry generators being: aware of meaningfulness, grammaticality and poeticness. Lamb, Brown, and Clarke (2017) propose a generic categorization of poetry generation methods that works on any creative task: mere, human-enhanced, and computer-enhanced approaches. Oliveira (2017) reviews poetry generation methods from multiple intersecting angles: languages, forms, content quality, techniques, and evaluation. In addition, Oliveira extends the list of techniques with: chart generation, constraint satisfaction, multi-agent, and language models such as Markov and RNN models. In this paper, we use a different taxonomy based on the intersection of chronology and methodology, pointing to the proposed creativity dimensions whenever applicable. We have found that the development of poetry generation techniques starts with rule-based, heuristic-based, statistical-based, then deep-learning-based approaches that are roughly proposed in this chronological order (Fig. 2).

[2] oed.com/view/Entry/146552

[3] merriam-webster.com/dictionary/poetry

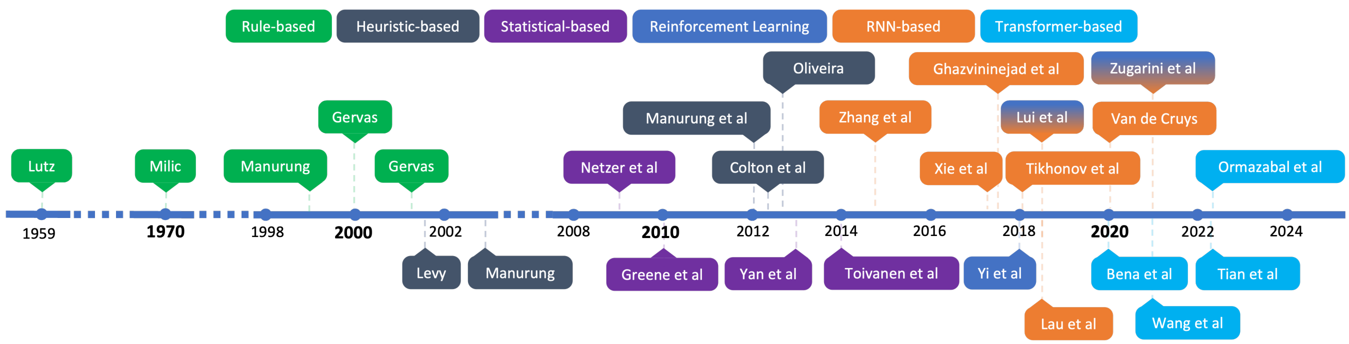


Figure 2: Poetry Generation Timeline.

Rule-based Approaches The early methods are based on following crafted rules and filling poem templates. Templates are the poem shells where actual words are taken out from original poems, while some of their characteristics, such as part of speech, are preserved as placeholders to be systematically filled with other words matching the characteristics. RETURNER (Milic 1970) is one of the earliest programs used to automatically generate English poems. The author designed three versions of a template-based model to generate three poems. The first version generates a poem given only the vocabulary of a real poem "Return by Alberta Turner." The second version was based on subjects and verbs with randomly decided optional modifiers, complements, and end conjunctions to provide the opportunity for iteration. The third version is based on a six-slot grammatical matrix and a set of rules to guide the generation process. Manurung (1999) uses a form-aware chart-based generation to generate strings that satisfy rhythm constraints. Meter validation is performed at each step of the process to reduce the search space. WASP (Gervás 2000) fills Spanish poem templates based on initial vocabulary and verse patterns, generating and validating verse drafts at each iteration according to length, rhyme, and rhythm constraints. The main focus of these early systems is on the poetic form rather than on the creative language. ASPERA (Gervás 2001) is an evolved version of WASP with case-based reasoning to retrieve an existing form and vocabulary that are highly similar to the user-defined target specifications and then adapt to meet the desired specifications. However, this model is not fully automated and needs to manually interact with a human user to validate or correct the generated draft.

Heuristic-based Approaches Using rule-based approaches is advantageous as they provide control and adherence to specific poetic forms or templates, but they are very predictable. Heuristic methods, however, are less predictable, and unpredictability is essential to creative writing. Levy (2001) proposes a poet-critic model theory that consists of a computer-based generator and an evaluator module. The author did not implement the described theory but proposed a prototype that utilizes the concept of evolution to create original outputs and uses classical neural networks as evaluators trained on real-world human evaluations. The evolutionary algorithm is a meta-heuristic approach in which the best population of poems is chosen by evaluator modules, modified by crossover operators and mutation operators and continuously evaluated until a satisfactory result is obtained. Levy uses such an approach to enhance poem features, words and rhymes. Manurung (2003) uses a similar approach with the following operations: add, delete, or edit as forms of mutation. In a later update of the framework, Manurung, Ritchie, and Thompson (2012) point to the fact that abusing the poetic license may lead to non-sense in a poetic form and that poetry generation approaches must conform to a restricted definition of poetry to generate falsifiable output. Similarly, Oliveira (2012) uses extracted semantic networks, grammar generators, and syllabic templates to generate Portuguese poems using genetic algorithms. Colton, Goodwin, and Veale (2012) address the lack of story behind the generated poems. The authors avoid to retrieve from a corpus of existing poems to restrain plagiarism, instead they choose short similes and non-poetic text corpora. They combine extracted key phrases from the non-poetic corpus with produced variations of existing similes to fill a user-given template. They also attempt to automatically generate a commentary on the generation process by recording context statements at each step of the generation.

Statistical Approaches A problem in heuristic-based models is that they are not capable of modeling the language. Netzer et al. (2009) uses statistical methods to analyze a corpus of English Haikus and generate words related to a theme provided by a user. The authors pick a real Haiku poem POS template at random and fill-in words from the theme set. The sentences are then ranked according to the degree of association of words. In order to augment the unpredictability of the output, they give more weight to pick second degree association instead of first degree. Greene, Bodrumlu, and Knight (2010) focus on the use of statistical methods to analyze, generate and translate poems. They use finite state transducers to map English sentences to a sequence of syllables and perform unsupervised learning to analyze the stress patterns of words in a poetry corpus. They used a finite state cascade to generate English love poems. In addition, the authors used a statistical phrase-based machine translation to translate Italian poems into English with the help of an iambic pentameter model. Poetry generation was also formulated as a text summarization task (Yan et al. 2013): Given the intentions of a user, write a poem that meets the poem requirements by retrieving terms from a poetry corpora. The authors use generative summarization and aim to produce correlated verses taking into consideration the poetic form and semantic coherence. Word associations were also used to fill poetry templates (Toivanen et al. 2014) replacing words from a poem with other words based on syntactic similarity. They use the templates of poems from the Imagist movement to emphasize figurative language.

Deep Learning Approaches Chomsky criticized statistical models as incomprehensible and inefficient compared to humans (Norvig 2017). The substantial amount of digital data created each second, the recent advances in computational power, and the ability of deep learning methods to derive complex data correlations, have led to state-of-the-art performance in NLG, including poetry generation.

• ***RNN-based*** RNNs were first introduced by John Hopfield in the early 1980s. However, they became popular as language models only a decade ago (Mikolov et al. 2010) for their ability to model long-term dependencies in sequential data such as text. Zhang and Lapata (2014) use a Convolutional Sentence Model (CSM) to capture sentence characteristics and use a character-based Recurrent Neural Network (RNN) model to incrementally generate poems. A set of keywords provided by the user is expanded with a poetic phrase taxonomy to generate candidate phrases, the top-ranked phrase will be the first line in the poem, and the following lines are sequentially generated given the previous line representations. The training process considers both the poetic form and semantics. Xie, Rastogi, and Chang (2017) designed two RNN-based models where the first combines char-level and word-level LSTMs with GloVe word embeddings for the model to capture not only the semantics of the words, but also morphemes and rhymes during training. The second model uses CNNs to generate word embeddings. Both models showed significant performance in coherence, poeticness, and form characteristics over vanilla char-based and word-based RNN models. Ghazvininejad et al. (2017) proposed a sonnet generation system based on RNNs and finite state acceptors and used embedding rhymes with a word2vec model. The system is interactive and allows the user to input topic words, adjust style configurations, and evaluate the generated poem with a 5-star rating; the evaluation was used to update the system performance. Lau et al. (2018) proposed a multi-model sonnet generator trained on Shakespeare's sonnets by means of bidirectional word-based and char-based LSTMs with the attention mechanism. The model consists of language, rhythm and rhyme submodels with Bi-LSTMs with an attention mechanism; the paper showed that Bi-LSTMs can capture poetic forms such as rhythm and rhyme very efficiently. Tikhonov and Yamshchikov (2018) used phoneme-based and char-based Bi-LSTMs, word-based LSTM, word embeddings and document embeddings for an author-stylized multilingual text generation framework with an application on English and Russian poems. Van de Cruys (2020) proposed an encoder-decoder model with double-layered GRUs using the attention mechanism. The decoder generated English and French poems line by line starting with the rhyming words backward to ensure a coherent verse with the forced rhyming scheme. To show the model's capability of presenting poeticness from scratch, the authors trained their model on non-poetic corpora. They proposed a latent semantic model to ensure topic coherence and a global optimization framework to score the verses and keep the highest-scoring ones.

• ***RNNs and reinforcement learning (RL)*** RNNs are able to handle input sequences of variable length, but they may suffer from gradient vanishing during backpropagation, which can cause the loss of distant context in the generated text. Additionally, Yi et al. (2018) argue that models based on maximum likelihood estimation, such as RNNs, tend to only optimize token-level loss rather than considering the poem as a whole, as humans do. To address this, they applied mutual RL to generate classic Chinese poems. The authors presented several reward modules instead of maximum likelihood to better mimic human behavior with reward schemes for fluency, coherence, meaning, and overall quality. Zugarini et al. (2021) used deep RL by combining Bi-LSTM models and RL in an iterative refinement approach to generate and revise poems following an author style and a rhyme scheme. Liu et al. (2018) also proposed a deep RL using a multi-adversarial CNN-RNN-based GAN model to caption images with a free verse English poem.

• ***Transformer-based*** Transformers (Vaswani et al. 2017) were introduced to overcome RNNs limitations caused by their sequential design, by allowing parallelization of training. Instead of processing a sequence token by token, sequences are processed holistically with positional encoding. Additionally, transformers introduced self-attention, allowing inputs-to-inputs and outputs-to-outputs attention to enhance context encoding and decoding. Bena and Kalita (2019) fine-tuned GPT-2 on a dream corpus to endow imagery language and fine-tuned it again on various emotion-based subcorpora to invoke one of the following emotions on the reader: joy, trust, anticipation, anger, fear, surprise, sadness, and disgust. LimGen (Wang et al. 2021) used GPT-2 to predict the next words in multi-candidate templates extracted from human-written limerick, then filtered out the words for meter and rhyme consistency, and output the top N lines using a variation of beam search that calculated a diversity score between templates. Ormazabal et al. (2022) proposed a structure-aware training of an autoregressive transformer model to generate formal verse poems in Spanish and Basque with only non-poetic corpora. Similarly, Tian and Peng (2022) proposed a multimodel framework trained on non-sonnet corpora to generate sonnets. The authors separated the training from the decoding and utilize a series of content planning, rhyme pair generation, polishing and constrained decoding to generate sonnets.

Recommendations

A major concern in rule-based and heuristic-based methods is the originality of language output, particularly when using poem templates and simple synonym replacements. Similar concerns apply to models trained on poetry corpora and those using transfer learning to limit mimicking the language, style, theme, and semantics of the source. Further study on active divergence in NLG can potentially enhance originality, enabling the production of creative content that incorporates novel elements. Another important area is optimizing decoding strategies through the use of enhanced sampling techniques, which can increase the unpredictability factor of the generated output. Recent advancements in improving nucleus sampling with randomized heads and introducing new sampling methods (Zhang et al. 2021b) have shown promising results in terms of diversity and novelty of generated texts. To address the sociability aspect of poetry, we recommend the utilization of iterative refinement models guided by discriminator models or reinforcement learning models to provide feedback during the generation process in a poet-critic-like framework. Additionally, we recommend developing creative LLMs that integrate elements of poetic diction during pretraining and incorporating multi-figurative generation models to aid in effective poetry generation.

Conclusion

In this work, we bring attention to the problem of generating creative data for researchers interested in the field. We analyze natural language generation models based on the three dimensions of creativity we proposed: originality, sociability and unpredictability. We provide a brief overview of text generation tasks and their potential creative applications, as well as a review of recent works. Finally, we provide a comprehensive overview examination of the task of poetry generation, including the methodologies and models employed in the literature. We hope that this review encourages researchers and provides insights for those interested in creativity in natural language generation and poetry, particularly in considering the use of suggested creativity and poeticness criteria in language models.

Acknowledgements

We thank Michael Ullyot for his insightful feedback and the anonymous reviewers for their valuable input.

References

Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; and Guibas, L. J. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11569–11579.

Bena, B., and Kalita, J. 2019. Introducing aspects of creativity in automatic poetry generation. In *Proceedings of the 16th International Conference on Natural Language Processing*, 26–35.

Berns, S., and Colton, S. 2020. Bridging generative deep learning and computational creativity. In *Proceedings of the Eleventh International Conference on Computational Creativity*, 406–409.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. Comet: Commonsense transformers for knowledge graph construction. In *Association for Computational Linguistics (ACL)*.

Brahman, F., and Chaturvedi, S. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5277–5294.

Brooks, J., and Youssef, A. 2021. Got: Testing for originality in natural language generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, 68–72.

Chakrabarty, T.; Ghosh, D.; Muresan, S.; and Peng, N. 2020. r^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7976–7986.

Chakrabarty, T.; Zhang, X.; Muresan, S.; and Peng, N. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4250–4261.

Chakrabarty, T.; Muresan, S.; and Peng, N. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6455–6469.

Chen, Y., and Lerch, A. 2020. Melody-conditioned lyrics generation with seqgans. In *2020 IEEE International Symposium on Multimedia (ISM)*, 189–196.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-face poetry generation. In *Proceedings of 3rd International Conference on Computational Creativity*, 95–102.

Colton, S.; Wiggins, G. A.; et al. 2012. Computational creativity: The final frontier? In *Ecai*, volume 12, 21–26. Montpelier.

Crimp, D. 1981. The end of painting. *October. Art World Follies (Spring, 1981)* 16:69–86.

Csikszentmihalyi, M. 1997. Flow and the psychology of discovery and invention. *HarperPerennial, New York* 39.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Garbacea, C., and Mei, Q. 2020. Neural language generation: Formulation, methods, and evaluation. *arXiv preprint arXiv:2007.15780*.

Gervás, P. 2000. Wasp: Evaluation of different strategies for the automatic generation of spanish verse. In *Proceedings of the AISB-00 symposium on creative & cultural aspects of AI*, 93–100.

Gervás, P. 2001. An expert system for the composition of formal spanish poetry. In *Applications and Innovations in Intelligent Systems VIII*. Springer. 19–32.

Gervás, P. 2002. Exploring quantitative evaluations of the creativity of automatic poets. In *Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science, 15th European Conference on Artificial Intelligence*.

Ghazvininejad, M.; Shi, X.; Priyadarshi, J.; and Knight, K. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, 43–48.

Goldfarb-Tarrant, S.; Feng, H.; and Peng, N. 2019. Plan, write, and revise: an interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 89–97.

Greene, E.; Bodrumlu, T.; and Knight, K. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 524–533.

Gu, J.; Wang, C.; and Zhao, J. 2019. Levenshtein transformer. *Advances in Neural Information Processing Systems* 32.

Hertzmann, A. 2018. Can computers create art? *Arts* 7(2):18.

Hockney, D. 2006. *Secret knowledge: Rediscovering the Lost Techniques of the Old Masters*. London: Thames & Hudson, new and expanded ed edition. OCLC: ocm70399577.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jahanian, A.; Chai, L.; and Isola, P. 2019. On the steerability of generative adversarial networks. *arXiv preprint arXiv:1907.07171*.

Keukeleire, P. 2020. Correspondence between perplexity scores and human evaluation of generated tv-show scripts. *M.S. thesis, Dept. Comput. Sci. Eng., Delft Univ. Technol., Delft, The Netherlands*.

Lai, H., and Nissim, M. 2022. Multi-figurative language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, 5939–5954.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2017. A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts* 11(3):159–179.

Lau, J. H.; Cohn, T.; Baldwin, T.; Brooke, J.; and Hammond, A. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1948–1958.

Levy, R. P. 2001. A computational model of poetic creativity with neural network as measure of adaptive fitness. In *Proceedings of the ICCBR-01 Workshop on Creative Systems*.

Liu, B.; Fu, J.; Kato, M. P.; and Yoshikawa, M. 2018. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM international conference on Multimedia*, 783–791.

Loller-Andersen, M., and Gambäck, B. 2018. Deep learning-based poetry generation given visual input. In *Proceedings of the Ninth International Conference on Computational Creativity*, 240–247.

Low, P. A. 2011. Translating jokes and puns. *Perspectives: Studies in Translatology* 19(1):59–70.

Lutz, T. 1959. Stochastische texte. *augenblick* 4(1):3–9.

Manin, D. 2021. Running in shackles: The information-theoretic paradoxes of poetry. In *Handbook of the Mathematics of the Arts and Sciences*. Springer. 1001–1014.

Manurung, R.; Ritchie, G.; and Thompson, H. 2012. Using genetic algorithms to create meaningful poetic text. *Journal of Experimental & Theoretical Artificial Intelligence* 24(1):43–64.

Manurung, H. M. 1999. A chart generator for rhythm patterned text. In *Proceedings of the First International Workshop on Literature in Cognition and Computer*, 15–19.

Manurung, H. 2003. An evolutionary algorithm approach to poetry generation. *Ph.D. thesis, University of Edinburgh*.

Markowitz, S. J. 1994. The distinction between art and craft. *Journal of Aesthetic Education* 28(1):55–70.

Mikolov, T.; Karafiát, M.; Burget, L.; Černockỳ, J.; and Khudanpur, S. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Milic, L. T. 1970. The possible usefulness of poetry generation. *The Computer in Literary and Linguistic Research, ed. R. A. Wisbey.* 169–82.

Mittal, A.; Tian, Y.; and Peng, N. 2022. Ambipun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1053–1062.

Mueller, J.; Gifford, D.; and Jaakkola, T. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, 2536–2544. PMLR.

Murfin, R. C., and Ray, S. M. 2009. *The Bedford Glossary of critical and literary terms*. Bedford/St. Martins.

Netzer, Y.; Gabay, D.; Goldberg, Y.; and Elhadad, M. 2009. Gaiku: Generating haiku with word associations norms. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, 32–39.

Norvig, P. 2017. On chomsky and the two cultures of statistical learning. *Berechenbarkeit der Welt?* 61–83.

Oliveira, H. 2009. Automatic generation of poetry: An overview. *Universidade de Coimbra*.

Oliveira, H. G. 2012. Poetryme: A versatile platform for poetry generation. *Computational Creativity, Concept Invention, and General Intelligence* 1:21.

Oliveira, H. G. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th international conference on natural language generation*, 11–20.

Oprea, S.; Wilson, S.; and Magdy, W. 2021. Chandler: An explainable sarcastic response generator. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 339–349.

Ormazabal, A.; Artetxe, M.; Agirrezabal, M.; Soroa, A.; and Agirre, E. 2022. Poelm: A meter-and rhyme-controllable language model for unsupervised poetry generation. *arXiv preprint arXiv:2205.12206*.

Paisley, W. J. 1966. The effects of authorship, topic, structure, and time of composition on letter redundancy in english texts. *Journal of verbal learning and verbal behavior* 5(1):28–34.

Preminger, A. 2016. *Princeton Encyclopaedia of Poetry and Poetics*. Macmillan International Higher Education.

Reiter, E. 2016. Ehud reiter's blog. `www.ehudreiter.com/2016/12/18/nlg-vs-templates/`. [Online; accessed 10-Jan-2023].

Riedl, M. 2020. Weird ai yankovic: Generating parody lyrics. *arXiv preprint arXiv:2009.12240*.

Shannon, C. E. 1951. Prediction and entropy of printed english. *Bell system technical journal* 30(1):50–64.

Shneiderman, B. 2007. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM* 50(12):20–32.

Steadman, P. 2002. *Vermeer's Camera: Uncovering the Truth Behind the Masterpieces*. Oxford University Press, USA.

Stowe, K.; Chakrabarty, T.; Peng, N.; Muresan, S.; and Gurevych, I. 2021. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

Teller, R. J. 2014. *Tim's Vermeer*. [Film]. Sony Pictures Classics.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Tian, Y., and Peng, N. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3587–3597.

Tian, Y.; krishna Sridhar, A.; and Peng, N. 2021. Hypogen: Hyperbole generation with commonsense and counterfactual knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1583–1593.

Tikhonov, A., and Yamshchikov, I. P. 2018. Guess who? multilingual approach for the automated generation of author-stylized poetry. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 787–794. IEEE.

Toivanen, J.; Gross, O.; Toivonen, H.; et al. 2014. The officer is taller than you, who race yourself!: Using document specific word associations in poetry generation. In *Proceedings of the Fifth International Conference on Computational Creativity*. Jožef Stefan Institute.

Tong, X.; Shutova, E.; and Lewis, M. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4673–4686.

Uehara, K.; Mori, Y.; Mukuta, Y.; and Harada, T. 2022. Vinter: Image narrative generation with emotion-arc-aware transformer. *arXiv preprint arXiv:2202.07305*.

Van de Cruys, T. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2471–2480.

Vaswani, A.; Shazeer, N.; Parmar, N.; G. A. N. Uszkoreit, J.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762,.

Wang, J.; Zhang, X.; Zhou, Y.; Suh, C.; and Rudin, C. 2021. There once was a really bad poet, it was automated but you didn't know it. *arXiv preprint arXiv:2103.03775*.

Xie, S.; Rastogi, R.; and Chang, M. 2017. Deep poetry: Word-level and character-level language models for shakespearean sonnet generation. In *Natural Lang. Process. Deep Learn.*

Xu, W., and Carpuat, M. 2021. Editor: an edit-based transformer with repositioning for neural machine translation with soft lexical constraints. *Transactions of the Association for Computational Linguistics* 9:311–328.

Yan, R.; Jiang, H.; Lapata, M.; Lin, S.-D.; Lv, X.; and Li, X. 2013. I, poet: automatic chinese poetry composition through a generative summarization framework under constrained optimization. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

Yi, X.; Sun, M.; Li, R.; and Li, W. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3143–3153.

Yin, Z.; Reuben, F.; Stepney, S.; and Collins, T. 2021. "a good algorithm does not steal–it imitates": The originality report as a means of measuring when a music generation algorithm copies too much. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*, 360–375. Springer.

Yu, Z.; Tan, J.; and Wan, X. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1660.

Zhang, X., and Lapata, M. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 670–680.

Zhang, Y., and Wan, X. 2022. Mover: Mask, over-generate and rank for hyperbole generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 6018–6030.

Zhang, J.; Cui, Z.; Xia, X.; Guo, Y.; Li, Y.; Wei, C.; and Cui, J. 2021a. Writing polishment with simile: Task, dataset and a neural approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 issue 16, 14383–14392.

Zhang, X.; Sun, M.; Liu, J.; and Li, X. 2021b. Lingxi: A diversity-aware chinese modern poetry generation system. *arXiv preprint arXiv:2108.12108*.

Zugarini, A.; Pasqualini, L.; Melacci, S.; and Maggini, M. 2021. Generate and revise: Reinforcement learning in neural poetry. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Are Language Models Unsupervised Multi-domain CC Systems?

Robert Morain, Branden Kinghorn and **Dan Ventura**
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
rmorain2@byu.edu, brando3034king@hotmail.com, ventura@cs.byu.edu

Abstract

Recently, ChatGPT has grown in popularity due to its ability to generate high quality text in a wide variety of contexts. In order to determine whether ChatGPT threatens to undermine the need for traditional CC systems, ChatGPT's ability to generate textual creative artifacts needs to be formally analysed. To do this, we constructed a survey that compares artifacts generated by traditional CC systems with corresponding artifacts generated by ChatGPT. Both types of artifacts are also evaluated independently on how well they possess certain desirable characteristics. Overall, the survey shows that artifacts generated by ChatGPT are preferred 36.84% ($p = 0.014$) more often and rated higher by 0.5 mean Likert scale points ($p = 0.0004$). These results indicate a need to reconsider the purpose and approach of traditional CC systems going forward.

Introduction

Computational creativity (CC) researchers often create applications that address creativity in specific domains such as stories (Pérez and Sharples 2001), poetry (Boggia et al. 2022), or puns (Ritchie 2003). These CC systems often introduce novel methods for generating creative artifacts such as templates, rules, or machine learning models. The authors then evaluate these generated artifacts either automatically or by way of a user survey. Recently, ChatGPT (OpenAI 2023) has demonstrated impressive text generation abilities. In this paper, we aim to evaluate ChatGPT's ability to generate creative artifacts by comparing ChatGPT's artifacts to artifacts generated by domain specific CC systems. While the scope of these experiments could include other modalities such as images (Ramesh et al. 2021), this paper focuses on textual creative artifacts.

This paper uses a definition of creativity that focuses on the generated artifact rather than on the process by which is created (Wiggins 2006).

Motivation

As statistical large language models improve, the need for domain-specific CC systems requires further consideration. If traditional CC systems are to remain relevant, they must offer distinct advantages over models like ChatGPT and its successors. ChatGPT implicitly learns many language related tasks through the general tasks of autoregressive language modeling (Radford et al. 2018) and fine-tuning with human feedback (Ouyang et al. 2022). The extent to which these abilities overlap with traditional CC systems is the central focus of this paper. Another topic related to this work is the philosophical question surrounding the creative limitations of language models.

All of the data used for this paper including prompts, survey results, and analysis is available online.[1]

Methods

The survey used in this paper includes artifacts from a selection of CC systems covering four textual domains: puns, six word stories, poetry, and short stories. These systems are JAPE (Ritchie 2003), PAUL BOT (Tyler, Wilsdon, and Bodily 2020), Movie Title Puns (Hämäläinen and Alnajjar 2019), MICROS (Spendlove, Zabriskie, and Ventura 2018), HIEROS (Spendlove and Ventura 2020), a GPT-2 model fine-tuned on six-word stories (Chien 2020), mBART for poetry generation (Boggia et al. 2022), TwitSong 3.0 (Lamb and Brown 2019), PoeTryMe (Oliveira and Alves 2016), MEXICA (Pérez and Sharples 2001), INES (Concepción, Gervás, and Méndez 2019), and Chronicle (Pickering and Jordanous 2017).

Notice that this survey does not discriminate against CC systems that use language models. The focus of the survey is not to determine whether language models are useful in CC systems, but rather to evaluate ChatGPT's creative text generation abilities compared to existing CC systems regardless of the methods used in the system.

While many potential systems could have been selected for the survey, these systems were chosen based on a combination of factors: the availability of artifacts present in the paper; the popularity of the system in the CC community; suitability to the domain; and recency of the paper's publication. Interested readers should refer to these papers for more details about their methods and implementation.

A set of twelve artifacts was compiled by selecting one artifact from each paper, to reduce the possibility of a single system under performing and thus skewing the results. Because the highest scoring artifact is not reported in the pa-

[1] https://tinyurl.com/s4k5pwry

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

CC system artifact:	How is a nice girl like a sugary bird? They are both sweet chicks.
Prompt:	Think like a punster. How is a nice girl like a sugary bird?
ChatGPT:	Here's a possible punning answer to your riddle: How is a nice girl like a sugary bird? They're both tweethearts!

Table 1: To create an artifact using ChatGPT with the same subject as a CC system artifact, a prompt with the appropriate domain and subject matter is provided to the model. The generated artifact is manually extracted. ChatGPT does not receive the original artifact in the prompt.

pers, the artifact expected to perform best according to the authors' subjective judgement was chosen.

Given the set of selected artifacts from CC systems, corresponding artifacts with similar subject matter were generated using ChatGPT. [2] ChatGPT was prompted to create an artifact from a particular domain (pun, six-word story, etc.) that included the same subject matter as the original artifact. Table 1 provides an example for how these artifacts were generated. This process facilitates the comparison of artifacts based on quality rather than other factors such as preference of subject. In some cases, when the generated artifact was too long or did not possess the correct subject matter, ChatGPT was iteratively prompted to generate a suitable artifact. Otherwise, the first artifact generated was selected. Artifacts were also screened for plagiarism by searching the web for exact copies.

Next, a survey was created to evaluate the artifacts based both on reviewers' preferences and characteristics used by various authors to evaluate the corresponding CC systems. To evaluate preferences, reviewers are asked to choose between a CC system artifact and the corresponding ChatGPT artifact, in a side by side comparison. Reviewers also had the option to mark "no preference". The reviewers were not made aware of which artifact came from a specialized CC system and which came from ChatGPT. To evaluate artifacts based on their characteristics, reviewers rated each artifact based on how well they possessed each characteristic on a Likert scale (1: strongly disagree, 2: somewhat disagree, 3: neither agree nor disagree, 4: somewhat agree, 5: strongly agree). For puns, the evaluation characteristics are "funny," and "surprising"; for six-word stories, "coherent" and "impactful"; for poems, "meaningful" and "emotional"; and for short stories, "entertaining" and "surprising". These characteristics were selected from the evaluation criteria used by the original authors to evaluate the CC systems. In addition, artifacts from all four domains are also rated on how creative they are perceived to be.

The survey was distributed online through Facebook, Instagram, Twitter, and Reddit. On Reddit, the survey was sent to the r/ArtificialIntelligence, r/MachineLearning, r/deeplearning, and r/ChatGPT subreddits. The survey does not ask for respondents to identify themselves or to rate their own knowledge of AI or CC; therefore it is unknown whether the reviewers are experts or not. The survey is randomized such that the questions and answers appear in random order.

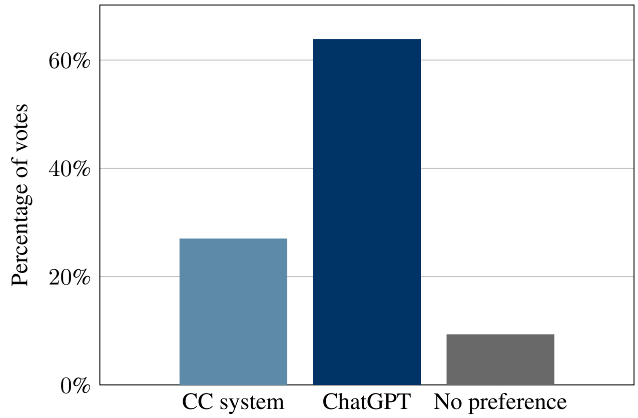


Figure 1: Reviewers' preferences in a one-to-one comparison between CC system generated artifacts and corresponding ChatGPT generated artifacts. These votes are aggregated across all domains and systems.

Results

Responses from 148 individuals resulted in an average of 39.5 responses to each question in the survey. Figure 1 shows reviewers' overall preferences across all domains and systems. The artifact produced by ChatGPT is preferred over the related CC system artifact 63% ($p = 0.014$)[3] of the time. However, the difference in terms of the characteristic evaluation of the two types of artifacts is relatively small. Figure 2 shows a difference of 0.50 Likert scale points ($p = 0.0004$), favoring the ChatGPT artifacts. Using the common significance threshold of 0.05, both of these results are statistically significant.

Figure 3 shows reviewers' preferences broken down by the four domains and aggregated across the three systems in each. For each domain, ChatGPT gains at least 61% of the votes. ChatGPT received the lowest percentage of votes in the poetry domain and the highest in the short story category with 77% of the votes. However, Figure 4 shows that the characteristic scores for the ChatGPT artifacts are relatively close to those for the original CC system artifacts. ChatGPT's lowest mean Likert scale score is in the pun domain with a score of 2.93 which is 0.10 points lower than the CC systems' score. The domain with the largest difference is the six-word story category with a margin of 0.62 points in favor of the ChatGPT artifacts.

The preferences for each artifact generated by their respective CC system along with the ChatGPT generated

[2]At the time of this experiment in April 2023, ChatGPT uses GPT 3.5 (See release notes: `https://help.openai.com/en/articles/6825453-chatgpt-release-notes`).

[3]Significance is calculated using a paired sample t-test.

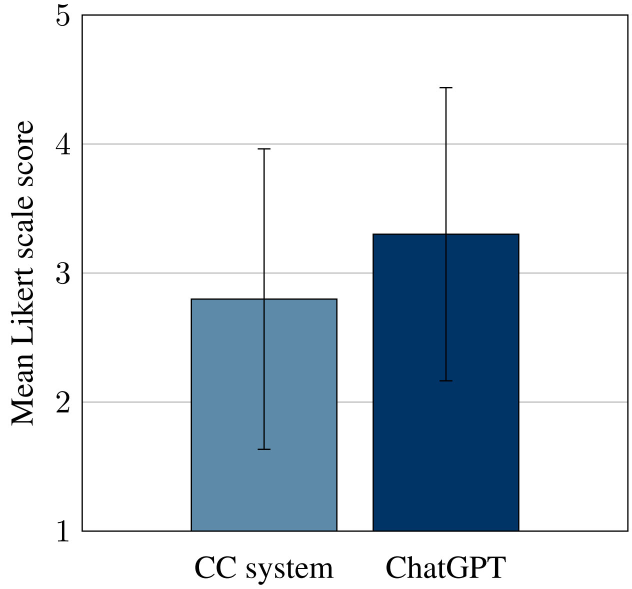


Figure 2: Characteristic evaluation of generated artifacts aggregated across all domains and systems.

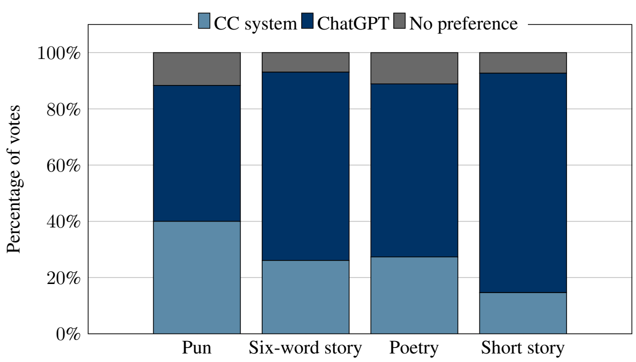


Figure 3: Reviewers' preferences aggregated across systems but broken down by domain.

counterpart is shown in Figure 5. For each system, the ChatGPT artifacts are preferred, with the exception of artifacts produced by PAUL BOT and Chien 2020. It is interesting to note that the INES system did not receive a single vote.

Figure 6 shows the mean Likert scale score for each artifact. The highest score overall belongs to (Chien 2020) which was generated by GPT-2 fine-tuned on a dataset of six-word stories. The characteristic evaluation scores usually correlate with the reviewers' preferences in that preferred artifacts have a higher score, with the exception of Chronicle which is preferred less but has a higher characteristic evaluation score than its ChatGPT counterpart.

For the characteristic evaluations, we can measure agreement between reviewers as a way to further assess our ability to be confident in the survey results, and this inter-rater agreement can be measured using Krippendorff's alpha (Krippendorff 2013). Across all systems and domains (cf.

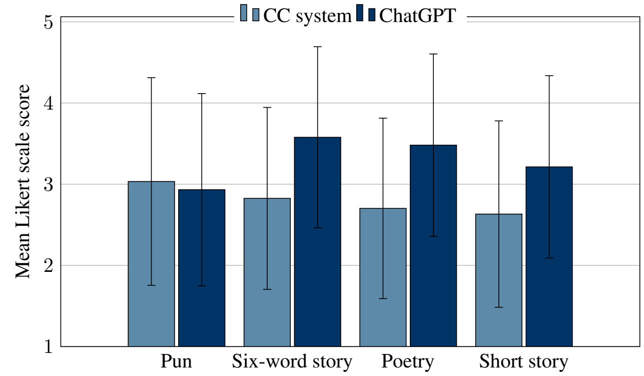


Figure 4: Reviewers' characteristic evaluation of artifacts in each domain, aggregated across systems.

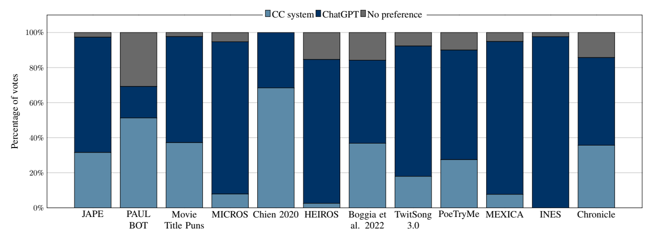


Figure 5: Reviewers' preferences broken down by the system that generated each artifact.

Fig 2), reviewer agreement produces $\alpha = 0.291$. Figures 7 and 8 show reviewer agreement broken down by domain (cf. Fig 4) and system (cf. Fig 6). Each of these values fall well below the recommended threshold of $\alpha \geq 0.8$ that would suggest reliable inter-rater agreement on preference for one system over another.

Discussion

The results seem to indicate that ChatGPT is able to generate artifacts that are just as good or better than the CC systems. This is similar to results found in (Radford et al. 2018) which shows that training a model on a general task like autoregressive language modeling leads to improved zero-shot performance on several downstream tasks as well. In this case, the data show that ChatGPT generalizes to creative tasks by outperforming CC systems overall, as well as at the domain and individual system level. The statistical significance of these results suggests that ChatGPT artifacts are likely to be preferred to and rated higher than (current/traditional) CC system artifacts.

While the results show that ChatGPT is capable of matching or surpassing CC systems in terms of the characteristic evaluation across all domains (Figure 4), the relative difference between CC system and ChatGPT artifacts is not as large as in the direct preferences analysis (Figure 3). In addition, the inter-rater agreement at the overall, domain, and system level is well below the recommended threshold for reviewer agreement, suggesting that characteristic evaluation does not completely explain reviewers' preferences.

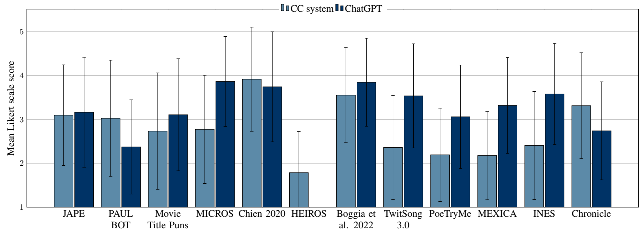


Figure 6: Reviewers' characteristic evaluation of each artifact.

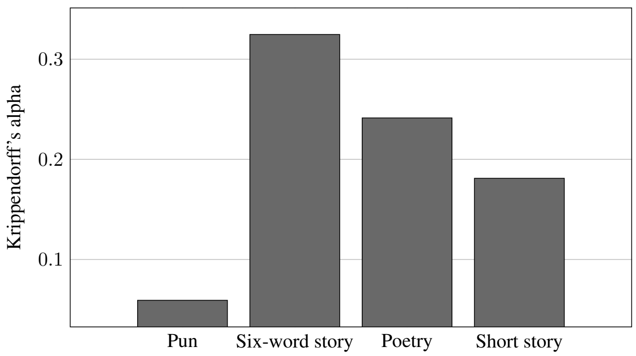


Figure 7: Agreement between reviewers by domain.

One reasonable explanation for this is that an artifact only has to be slightly better in order to be preferred. Although, the presence of a "no preference" option provides confidence that there is a real difference in preference between the artifacts, even if that preference is small.

It is also possible that the criteria used in the characteristic evaluation fail to capture all of the reasons why reviewers prefer an artifact. For example, large language models like ChatGPT are very capable of generating fluent text even if the content of the text is nonsense. In addition, there may be other positive characteristics that ChatGPT includes in its artifacts, such as accessibility to a general audience or even other domain specific characteristics.

It is also reasonable to conclude that the characteristic evaluation is reliable—reviewers generally prefer the ChatGPT artifacts, and while the difference between the artifacts in terms of their character evaluation is not large, the significance testing provides confidence that this difference is, in fact, real. Also, it is important to remember that the artifacts selected for the survey that came from CC systems are (presumably) the best those systems have to offer. On the other hand, ChatGPT's artifacts are not cherry picked and most of the artifacts were generated with a single non-engineered prompt. Therefore, it may be argued that these results may represent a comparison of the floor of ChatGPT's abilities to the ceiling of (traditional) CC systems' abilities.

Implications and Future Work

The findings of this survey do not discount the work of CC researchers. Rather, their accomplishments with significantly fewer resources indicate that many of these traditional

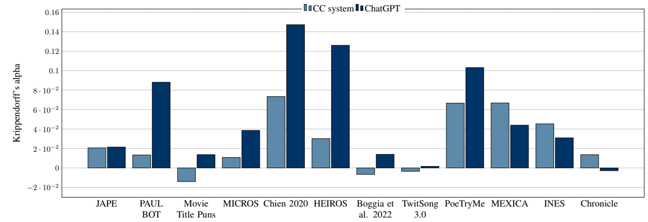


Figure 8: Agreement between reviewers by system.

CC systems are truly ahead of their time. It is also possible that the methods demonstrated by these systems applied at the scale of ChatGPT may outperform ChatGPT.

The purpose of this paper is to spark debate about the creative limitations of language models like ChatGPT and CC systems in general. Given that this level of performance comes from a general language model like ChatGPT means that the purpose and approach of domain-specific CC systems needs to be carefully considered. At the very least, ChatGPT should be used as a baseline when evaluating CC systems going forward.

ChatGPT represents a paradigm shift in terms of interactivity in creative systems. In these experiments, interactive prompts serve to constrain the system to produce corresponding artifacts that are comparable to their CC system counterparts. ChatGPT's ability to do this successfully demonstrates the system's robustness and ease of use. It also suggests a possible move away from fully autonomous systems towards more co-creative solutions (though this certainly doesn't preclude fully autonomous systems in any way, of course.)

These results also highlight an opportunity to improve the performance of language models on creative tasks. While the ChatGPT artifacts are preferred, the overall characteristic evaluation shows that reviewers still have a generally neutral attitude toward the artifacts. It is not yet clear from where these improvements will come, but it is possible that some help may be found in traditional CC approaches.

References

Boggia, M.; Ivanova, S.; Linkola, S.; Kantosalo, A.; and Toivonen, H. 2022. One line at a time — generation and internal evaluation of interactive poetry. In *Proceedings of the International Conference on Computational Creativity*, 7–11. Association for Computational Creativity.

Chien, G. 2020. Generating six-word stories. `http://cs230.stanford.edu/projects_fall_2020/reports/55790134.pdf`.

Concepción, E.; Gervás, P.; and Méndez, G. 2019. Evolving the INES story generation system: From single to multiple plot lines. In *Proceedings of the International Conference on Computational Creativity*, 220–227. Association for Computational Creativity.

Hämäläinen, M., and Alnajjar, K. 2019. Modelling the socialization of creative agents in a master-apprentice setting: The case of movie title puns. In *Proceedings of the Inter-*

national Conference on Computational Creativity, 266–273. Association for Computational Creativity.

Krippendorff, K. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, 3rd edition.

Lamb, C., and Brown, D. G. 2019. Twitsong 3.0: Towards semantic revisions in computational poetry. In *Proceedings of the International Conference on Computational Creativity*, 212–219. Association for Computational Creativity.

Oliveira, H. G., and Alves, A. O. 2016. Poetry from concept maps–yet another adaptation of PoeTryMe's flexible architecture. In *Proceedings of the International Conference on Computational Creativity*, 246–253. Sony Computer Science Laboratories.

OpenAI. 2023. Introducing ChatGPT. `https://openai.com/blog/chatgpt`. Accessed 2023-4-11.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

Pérez, R. P. Y., and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Pickering, T., and Jordanous, A. 2017. Applying narrative theory to aid unexpectedness in a story generation system. In *Proceedings of the International Conference on Computational Creativity*, 213–220. Association for Computational Creativity.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2018. Language models are unsupervised multitask learners. `https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf`. Accessed 2023-4-11.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8821–8831.

Ritchie, G. 2003. The JAPE riddle generator: technical specification. Technical Report EDI-INF-RR-0158, School of Informatics, University of Edinburgh.

Spendlove, B., and Ventura, D. 2020. Creating six-word stories via inferred linguistic and semantic formats. In *Proceedings of the International Conference on Computational Creativity*, 123–130. Association for Computational Creativity.

Spendlove, B.; Zabriskie, N.; and Ventura, D. 2018. An HBPL-based approach to the creation of six-word stories. In *Proceedings of the International Conference on Computational Creativity*, 161–168. Association for Computational Creativity.

Tyler, B.; Wilsdon, K.; and Bodily, P. 2020. Computational humor: Automated pun generation. In *Proceedings of the International Conference on Computational Creativity*, 181–184. Association for Computational Creativity.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge Based Systems* 19(7):449–458.

Tweetorial Hooks: Generative AI Tools to Motivate Science on Social Media

Tao Long[§♔], Dorothy Zhang[♕], Grace Li[♕], Batool Taraif[♕], Samia Menon[♔], Kynnedy Simone Smith[♔], Sitong Wang[♔], Katy Ilonka Gero[♔], Lydia B. Chilton[§♔]
[♔]Columbia University, New York, NY [♕]Barnard College, New York, NY
[§]{long, chilton}@cs.columbia.edu

Abstract

Communicating science and technology is essential for the public to understand and engage in a rapidly changing world. Tweetorials are an emerging phenomenon where experts explain STEM topics on social media in creative and engaging ways. However, STEM experts struggle to write an engaging "hook" in the first tweet that captures the reader's attention. We propose methods to use large language models (LLMs) to help users scaffold their process of writing a relatable hook for complex scientific topics. We demonstrate that LLMs can help writers find everyday experiences that are relatable and interesting to the public, avoid jargon, and spark curiosity. Our evaluation shows that the system reduces cognitive load and helps people write better hooks. Lastly, we discuss the importance of interactivity with LLMs to preserve the correctness, effectiveness, and authenticity of the writing.

Introduction

Communicating science and technology is important for the public to understand and engage in a rapidly changing world. Recently, a majority of the public learns about the world not from traditional publications, but from social media platforms (Shearer and Matsa 2018). *Tweetorials* are an emerging format for explaining complex scientific concepts on Twitter. They consist of a series of tweets that explain a technical concept in informal, narrative-driven ways (Breu 2019; 2020). Whereas typical science writing is often formal, the norms of social media allow scientific conversations to take on a more personal style (Brüggemann, Lörcher, and Walter 2020), allowing for more creative forms of expression and engagement.

The most important part of a Tweetorial is the first tweet. This is often called a "hook" because it aims to hook the readers' attention and spark their curiosity so they want to read more. Although there are many ways to do this, an analysis of Tweetorial hooks (Gero et al. 2021) has shown that a common pattern is to start with a specific, relatable experience that uses no jargon. However, the challenge is to find a common experience for technical topics that a general audience of readers will find engaging.

Many STEM experts want to write creative and engaging science-related content for the public, but are not trained to do so. Their writing training is mainly for writing to peers — other experts who are familiar with the motivation for the work, who expect expert terminologies, and who know the context surrounding the science and the formal culture of academic writing (Aldous, An, and Jansen 2019). Such writing is typically (and purposefully) formulaic, and creative writing may even be discouraged in such contexts. Although there are many theories, examples, and books about public science communication, they lack mechanistic strategies proven to help people use them (Howell et al. 2019; McClain and Neeley 2014; Yeo 2015). Providing explicit support for informal science writing like Tweetorial hooks can better support experts in writing for the public.

We explore various ways for large language models (LLMs) to help people write engaging, creative hooks for computer science topics. We first explore how well LLMs can write hooks on their own by investigating three prompting strategies: instructions, instructions and examples, instructions, examples, and relatable experiences. We find that although adding examples and experiences in the prompt improves hooks, the LLMs still have much room for improvement. Then, we design an interactive system that scaffolds the process of writing hooks but allows users to accept, reject, or improve LLM suggestions at every step. In a user study with ten people proficient in their domain and familiar with Tweetorial hooks, we show this drastically improves their hooks and reduces cognitive load compared to writing without the system.

Related Work on LLMs and Writing Support

Advances in LLMs have resulted in machine abilities to complete prompts with rich knowledge, commonsense reasoning, and fluent language composition (Radford et al. 2019). Despite not being explicitly trained for specific tasks, these models possess impressive generative capabilities and can perform a diverse range of tasks. Moreover, providing just a few examples in the prompt itself can significantly enhance the quality of the model's outputs (Brown et al. 2020).

LLMs show great promise for supporting creativity and writing tasks. They can help with story writing (Calderwood, Wardrip-Fruin, and Mateas 2022; Chung et al. 2022), brainstorming (Singh et al. 2022), and finding creative connections (Wang et al. 2023) as well as story angles from press releases (Petridis et al. 2023). They have been shown to help with all three stages of the cognitive process model of writing (Flower and Hayes 1981): planning/ideation, translating/drafting, and reviewing (Gero, Liu, and Chilton 2022). Rather than executing these stages in a linear fashion, the writing process typically involves iterative use of these stages and requires writers to switch between their writing goals while keeping their audience in mind (Emig 1977). Because of this, writing can be taxing on both the writer's short- and long-term memory, resulting in high cognitive demands (Hayes 1996). Thus, LLMs as a writing companion and support can benefit writers in reducing mental load.

Despite the successes of LLMs, problems remain. Language models tend to output repetitive and vague responses (Holtzman et al. 2020; Ippolito et al. 2019), particularly when a prompt is underspecified or too difficult to address. One approach to address this is to chain LLM prompts together (Wu, Terry, and Cai 2022): breaking down a problem into simpler and more explicit steps can make it easier for LLMs to complete. A bigger challenge is that language models have no model of truth. They learn correlations from large amounts of text, but they are not able to tell if the text they produce that includes falsehoods or offensive language (Bender et al. 2021). Thus, LLMs may best assist writers in producing higher-quality written outputs by providing support during the writing process instead of replacing the writer and writing on their own.

Headline writing is an established challenge in natural language processing. Fully automated systems have some successes at generating headlines (Bukhtiyarov and Gusev 2020), and some can even write ones in a "clickbait" style to hook readers (Jin et al. 2020; Xu et al. 2019). Although headlines do serve as hooks, traditional journalistic headlines have a different style than Tweetorial headlines. Tweetorial hooks are a little longer than headlines and use that space to start an engaging, relatable, and vivid personal anecdote. Thus the narrative, rather than the keywords, is the basis for engaging readers. This paper extends works on engaging readers with intriguing and meaningful content.

Background on Tweetorial Hooks

Tweetorials are a "collection of threaded tweets aimed at teaching users who engage with them" (Breu 2019). Across a wide range of topics from medicine, to climate science, to physics, to computer science, these tweets always introduce a technically complicated concept or answer a popular science question through informal, narrative-driven, and creative writing (Breu 2020; Gero, Liu, and Chilton 2022). Figure 1 shows the first, last, and some middle tweets that form the overall narrative. Hooks are the first tweet that grab readers' attention and pull them into the narrative.

Previous work (Gero et al. 2021) has analyzed Tweetorial hooks and described attributes of high-quality ones: 1) a relatable and interesting example as a lead-in and 2) an intriguing question that is driving and specific that sparks readers' curiosity. Relatable and specific content can take many forms. It can relate a topic to things in the news, refute a popular misconception, or take a common daily experience and help explain it. For the language to be relatable, the hook should not include jargon. Using unfamiliar technical terminology undermines the purpose of engaging the public (Bullock et al. 2019). Then, an intriguing question will be directly or implicitly proposed to the readers to help spark curiosity and draw them to the following thread. The unanswered question will connect the previous relatable example to the following threads of explanations. Thus, we establish a list of requirements (R) for a relatable and engaging Tweetorial hook:

- ***R1*** *- Jargon-Free:* Does the hook avoid jargon or unexplained terminology so the general audience can understand it easily?
- ***R2*** *- Specific and Relatable Example(s):* Does the hook include specific and relatable example(s) about the topic?
- ***R3*** *- Sparks Curiosity:* Does the hook drive readers to continue reading and satisfy their curiosity?

Here are two examples of Tweetorial hooks for computer science topics that exhibit all these properties:

- Virtual Private Network (VPN): *"I once torrented Last Week Tonight — then my landlord got a complaint from Comcast! WTH? My friends never got caught. Ugh. So here are things I wished I had known about how to be sneaky on the internet:"*

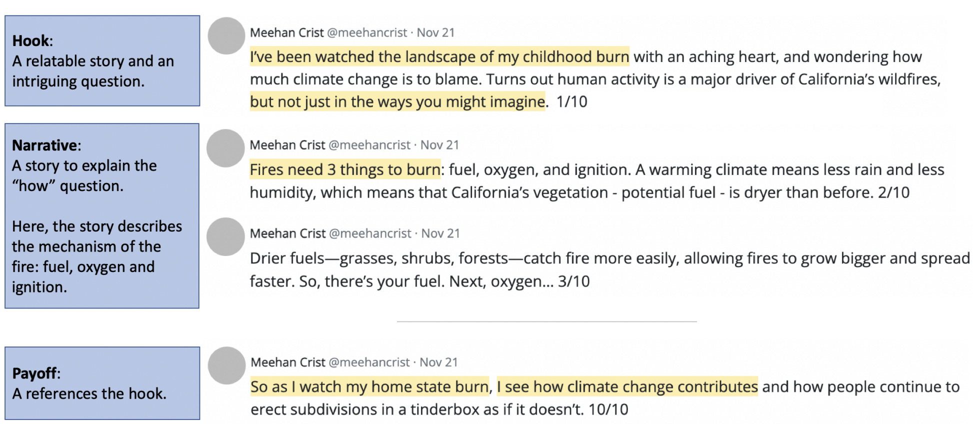


Figure 1: A Tweetorial about the California wildfires (Crist 2019) annotated for narrative structure. Yellow highlights indicate key phrases of the hook (including the relatable detail and the intriguing question), narrative, and payoff. More annotated Tweetorial examples can be found on our website: `http://language-play.com/tech-tweets/annotations`

- Language Models: *"My son relies on his Alexa to help with his math homework every single night. While I am concerned about his learning, I am interested in how Alexa understands what he is saying? Is it the same way that humans understand language? What is the difference? A thread on how language models helps with this:"*

For each hook, the topic is motivated by an everyday experience. For VPNs, the experience is torrenting. For language models, the experience is Alexa. Each experience is told in a personal way (*"then my landlord got a complaint from Comcast!"*), with informal language ("*Wth? My friends never got caught. Ugh.*"). Often they have very specific details (*"Last Week Tonight"*). They don't contain jargon—other than when mentioning the name of the topic towards the end of the hook. And they have a question or implied question that sparks curiosity and drives the reader to learn more (*"how to be sneaky on the internet"*). This is a lot to achieve in one tweet.

Studies on Tweetorial writing have shown that writing hooks is a key challenge for STEM experts (Gero et al. 2021). They are trained to write about their work in a formal tone for other experts, and it is difficult to go against that training. Also, they feel uncomfortable using subjective and informal language and avoid personal details, even though 80% of the Tweetorials have them.

In an exploratory study using LLMs to support Tweetorial writers, one of the major use cases was ideating concrete examples for the hook (Gero, Liu, and Chilton 2022). This indicates there is potential to help STEM experts write in informal styles. We build on this potential by studying LLMs' potential to write hooks, then designing a workflow to scaffold writers' hook writing process and using LLMs to suggest options for relatable experiences that are jargon-free and can spark curiosity.

Study 1: Prompt Engineering Study

We first investigate how well an LLM can write hooks without human intervention. Then, we compare the performance of three prompting strategies and use expert annotators to evaluate the outputs.

Participants and Procedures We identified 30 technical computer science topics that are important for a general audience to understand. We selected them randomly from the *Sideways Dictionary* (Jigsaw 2017) — a website for journalists to find accessible explanations for common technical terms. These terms included such as *Database*, *Browser Hijacking*, *Programming Language*, *Internet Service Provider*, and *Autocomplete* (See Appendix for the complete list).

The three prompting strategies (PS) we compared are:

- ***PS1** (Instructions only)* is the most basic strategy which asks for a hook and provides simple instructions that the hook should be jargon-free, include a relatable and specific example, and spark curiosity. This is the bare minimum needed to explain to the LLM the goal of a hook.
- ***PS2** (Examples and Instructions)* has all the instructions from PS1, and adds five examples of good hooks we identified and collected inside the team. These hooks were taken from published Tweetorials and edited lightly for clarity. Adding examples is a known technique to help the LLM learn the "styles" that are difficult to describe or to phrase in specific instructions such as writing objective, writing structure, diction, and tone.
- ***PS3** (Examples, Chained User Details, and Instructions)* is a three-stage pipeline that chains LLM prompts together (Wu, Terry, and Cai 2022), in addition to all the content from PS2. It first asks for the user's topic to generate everyday examples, then common experiences, then a specific personal anecdote about this experience. LLM chaining is known to work well when instructions are complicated. It breaks down the problem into simpler steps and builds up to a complex output.

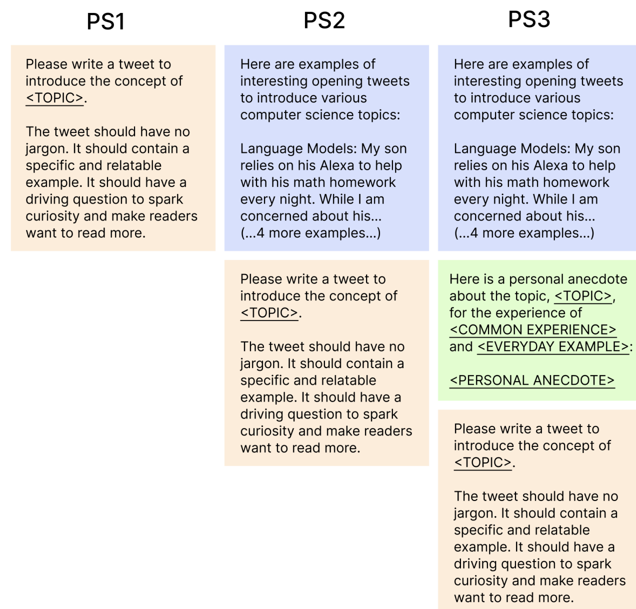


Figure 2: An illustration of the three prompting strategies

We used OpenAI's GPT-3 API and its *text-davinci-003* model with the default settings for all parameters, as it was the most capable model available at the time of our study.

In this study, we investigate the following hypothesis:

Hypothesis #1: PS3 will attain the highest overall score and outperform both PS1 and PS2 across all three rubric categories. We believe that the prompt chaining will break down the complex hook writing task into simpler steps that LLMs will be better able to solve one at a time.

To evaluate the three prompting strategies, we hired three annotators with professional training in communication and writing to judge the hooks' quality. Each annotator rated 270 hooks — 30 topics with three prompting strategies and three generations each. The annotators were paid $20 per hour and evaluated each hook on a 1 to 5 scale based on the criteria: whether it is jargon-free (R1), contains a relatable example (R2), and sparks curiosity (R3). They received a detailed annotation rubric with examples (See Appendix).

Results

Overall, the annotators had fair agreement on their assessment, with a Fleiss' kappa of 0.23.

According to our annotation results (See Figure 3), PS1 was the lowest-scoring strategy, with an average of 2.93 out of 5. PS2 and PS3 were only about half a point higher than PS1 at 3.49 and 3.47 out of 5, but about equal to each other. All three strategies performed pretty well at being jargon-free, even PS1. Seemingly, LLMs can follow the instruction to be jargon-free without examples. However, where PS1 struggled was in being relatable and sparking curiosity. Here, PS2 and PS3 performed 1 point better on relatability and almost 1 point better on curiosity. This indicated that the training examples in PS2 and PS3 did help LLMs "learn" how to write a more relatable hook with details.

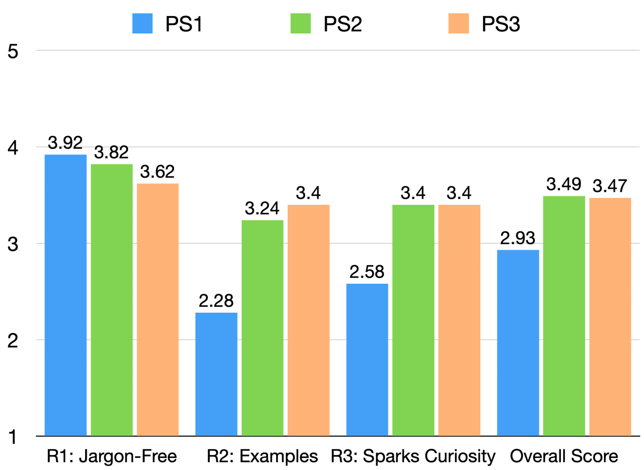


Figure 3: Average scores for each prompting strategy based on rubric performance

To answer our Hypothesis #1, **PS3 and PS2 were similarly good, and both were better than PS1.** Specifically, PS3 was only significantly better than PS1 for R2 (p-value $<$ 0.001), R3 (p-value $<$ 0.001), and the overall performance (p-value $<$ 0.001). However, compared to PS2, PS3 performed similarly to PS2 in all categories. This was surprising because the average score of PS2 (SCORE - 3.49/5) left much room for improvement. We hoped the chaining in PS3 would improve the hook quality, but it did not.

One reason for PS3's unideal performance was that, PS3 often included jargon and failed to be relatable, though PS3 provided more detailed experiences. For example, the lowest-scoring hook from PS3 on Table 4, we saw that, with a topic of Back End, it did not give out a more detailed experience than what PS2 usually had: "*my recent experiences with Amazon Web Services' Identity and Access Management feature...*" It reflected a problem that PS3 often included details that were specific but not relatable and even contained jargon or unexplained terms like "*Amazon Web Services*," "*Identity and Access Management*," and "*bad end access*." Clearly, this experience was not relatable to general audiences, though it was detailed. Thus, for the lack of improvement from PS2 to PS3, we can see the lack of manual filtering of the specific. However, with humans in the loop, the process of picking better answers would help improve answers at every step and make the final results closer to the rubric. Thus, to understand whether human interventions help with the PS3, we conducted the following study.

Study 2: User Study

We conducted a user study to evaluate the effectiveness of our LLM-based Tweetorial solution for users with the need to communicate science to the general public.

System Description We built an interactive web application using HTML, Python, Javascript, Flask, and the GPT-3 API to help users write engaging hooks for technical topics. The interface scaffolded the process of writing a hook into steps and used GPT to generate suggestions that the users can regenerate, modify, or accept before going to the next stage. The system and the workflow can be seen in Figure 4:

- *Step 1*: Users write down their topic, and the system generates five concrete everyday examples.
- *Step 2*: Users input their everyday example (either from the previous step or their own answer), and the system generates five common experiences people might have with that example.
- *Step 3*: Users input a common experience, and the system generates three separate personal anecdotes.
- *Step 4*: Users feed their favorite personal anecdote back to the system to add details to make it specific and vivid.
- *Step 5*: Users input their final anecdotes, and the system generates an example hook based on all previous inputs.
- *Step 6*: Users write a final hook by either directly taking the LLM-generated one or adjusting it accordingly.

Participants and Procedures We recruited ten participants from a local college student network and asked them to write Tweetorial hooks with and without our prototype in February 2023. The participants included six females and four males, with an average age of 20.1 years old. All ten users had expertise in computer science and familiarity with the particular topics we were asking them to write about. The study took around 1.5 hours, and they were paid $30.

Before the study, participants first received a 10-minute introduction to Tweetorials and hooks. The introduction included explanations and examples of what constitutes a good hook. Then, they were asked to write hooks for six randomly chosen computer science topics from the list we used for the annotation study. The topics, in sequential order, were *Front End*, *Autocomplete*, *Programming Language*, *Net Neutrality*, *Application Programming Interface (API)*, and *Cybercrime*. For the hook-writing tasks, we asked each of them to write on three topics using the system and three without the system. The participants were randomly assigned to two groups, each consisting of five. Group 1 wrote with the system for the first, third, and fifth topics, and without the system for the others. Group 2 followed the opposite order. This approach ensured a fair comparison by evenly distributing the system use across all topics and participants.

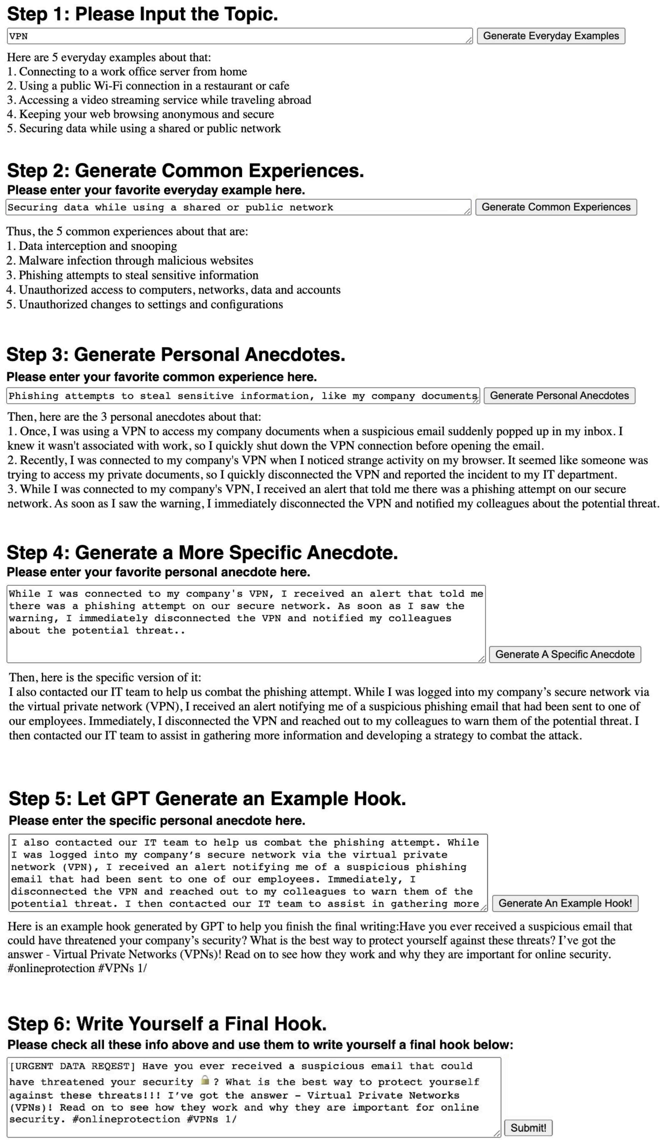


Figure 4: An example of how users write a Tweetorial hook about *VPN (Virtual Private Network)* with our tool. They can follow the workflow from top to bottom or return to previous steps and start again. They can regenerate, modify, and accept the LLM outputs or use their own responses.

During each hook-writing task, we first provided the participants with the topics and informed them whether to use the system. Then, they had eight minutes to write a hook. During the session, users were informed that they could search for information online regardless of the conditions. After each hook writing task, we asked them to fill out a NASA Task Load Index (TLX) (Hart and Staveland 1988) questionnaire to understand their mental load and experiences quantitatively. After finishing all six writing tasks, we started a 25-minute semi-structured interview to learn more about their experiences and hook writing process.

In this study, we investigated the following hypothesis:

Hypothesis #2: Using the system reduces the mental load and increases the performance of writing hooks.

Results

The TLX results are visualized in Figure 5 and Table 1. As we split participants into two groups for randomization, they had good internal consistencies within each group, with Cronbach's Alphas of 0.78 and 0.85.

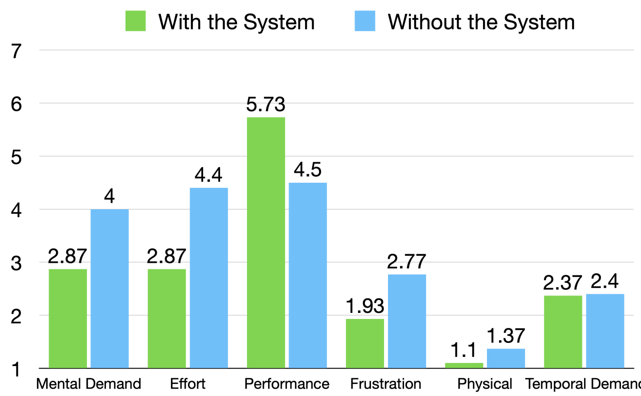


Figure 5: User study TLX results

TLX Dimension	With System	Without	p-value
Mental Demand	2.87	4.00	**0.004****
Effort	2.87	4.40	**0.002****
Performance	5.73	4.50	**0.001****
Frustration	1.93	2.77	**0.02***
Physical Demand	1.10	1.37	0.08
Temporal Demand	2.37	2.40	0.598

Table 1: User study TLX results and p-values for Wilcoxon tests (** indicates statistical significance at the p<.005 level, * indicates statistical significance at the p<.05 level)

1. Less Mentally Demanding The TLX scores indicated that writing hooks was less mentally demanding with the system (SCORE - 2.87/7) than without it (SCORE - 4.00/7, p-value = 0.004). All ten users expressed that without the system, it was hard to find concrete and specific examples of abstract topics. Under that condition, many users did their own brainstorming, often trying to think of their own experiences with the topic and attempting to recall tangible details and emotions about it before they were able to start writing (P8). Five users said that even if they did come up with a few examples, it was challenging to narrow them down to one to fit the criteria: relevant, relatable, and interesting enough to make them keep reading (P1, P2, P5, P7, P8).

All ten users expressed the ease of using the system to help simplify language into digestible terms that more people can understand. P4 shared it is easier to brainstorm a lot of ideas, and it helped open horizons and applications, but they still ended up choosing one that resonated the most. P1, P2, P5, P7, P8, and P10 mentioned that the workflow was straightforward, easy, clear, and simple to use, easing mental burdens during the hook writing process. All ten users said they would use this tool in the future.

2. Less Effort The TLX scores indicated that writing hooks required less effort with the system (SCORE - 2.87/7) than without it (SCORE - 4.40/7, p-value = 0.002). Under the

Topic	Without the System	With the System
Application Programming Interface (API)	Have you heard about the huge Oracle and Google lawsuit but had no idea what it was about? What the hell even is an API, and why is it so important that they can't be copyrighted? A guide for Supreme Court Justices (and you)	Ever wonder why your Spotify Wrapped is so fun? How do they know which artists and songs to highlight and recommend? Find out how Spotify's developer tools can help analyze user listening history and trends to make tailored music content in this thread:
Cybercrime	These days computers are a huge part of our lives- what illegal activities could be going on within our computers? In this thread, we will be exploring cybercrime, and what this could mean for our online safety. 1/	Have you ever received a call out of the blue from someone claiming to be from your bank, asking for your personal information? After this happened to me recently, I wondered what other kinds of cybercrime exist and how someone like me can protect themselves? Here's what I found out:

Table 2: Collection of hooks generated under both the with-system and without-system conditions from the user study

without system condition, seven users spent a lot of effort searching the Internet to find examples without much success. Even though there were some examples on Google, it was hard and time-consuming for users to find them. P2 and P5 shared that Google felt like an "*ocean of information*." They had to spend a long time searching: skimming through the titles, avoiding getting technical information, and clicking on it to understand the material first and then adapting it to their own work. They needed to put down three to five search queries on Google to find the results they wanted. For example, P9 used "*net neutrality examples*" and "*net neutrality in simple terms*"; P2 used "*examples of APIs we use in our everyday lives*", "*define programming language in a fun way*," "*explain the term front end for a 5-year-old*" and "*what is the front end for dummies*"; P7 used "*examples of popularly used APIs*" and "*how to talk about programming languages in layman's terms*." Trying different terms took a long time and effort (P8) and often ended in failure (P1).

In contrast, P1, P2, P4, P6, and P7 all mentioned that the with-system experience was just effortless: "*easy to generate and regenerate*", "*easy to find strong ideas*", and easily "*reminded me of what I already knew*". P8 shared that the writing workflow was seamless, enabling them to complete the hook writing process by following the steps without searching on Google. In total, eight of the ten participants finished the with-system writing process without Google.

3. Better Performance The TLX scores indicated that users achieved better performance writing hooks with the system (SCORE - 5.73/7) than without it (SCORE - 4.50/7, p-value = 0.001). Users felt more confident and satisfied with the results they obtained from the system when using LLMs, as they believed that the process involved fewer personal biases and LLMs had more knowledge about real common experiences. For instance, P8 mentioned that they believed the common experiences generated by LLMs were meant to be more familiar and relatable to the general public. In comparison, they reported concerns that the experiences they came up with on their own or from Google were not common enough and biased toward their personal background. Similarly, P2, P4, and P7 shared that they experienced these implicit biases and received fewer affirmations while trying to write a hook without LLMs, as they trusted LLMs more.

4. Users Edit LLM Hooks to Meet Requirements In Step 5 of the system, users were presented with a hook written by the LLM based on their responses to Steps 1-4. All ten users expressed that the LLM-generated hooks are good and useful, while six of them expressed the need to edit the LLM-generated hooks to make them more relatable and engaging. When asked to make a quick comparison between their edited version and the LLM-generated ones, all these six writers shared that their edits were necessary and helped elevate the quality of the hooks.

Responding to R1 (being jargon-free), P1, P8, and P10 shared that they still found jargon inside the LLM-generated hooks. Thus, they removed the unexplained terminology or hard-to-understand acronyms. For example, P10 replaced the acronym of "*ISPs*" from the LLM-outputted hook with "*Internet Service Providers*." They had concerns that the system might overlook requirements after chaining too much stuff. Also, they edited the hook for conciseness by cutting off extra questions and wordy introductions.

For R2 (including relatable and specific examples), several writers said that the LLM output felt robotic and rigid, thus making it less engaging (P1, P2, P5). For example, P1 mentioned that when they read the LLM-generated hook, they felt it would not interest readers. Also, P2 shared that the first sentences in many LLM-generated hooks felt like news headlines, which read like some emotionless statements. Thus, they edited the tone to become funnier and more personable. Also, P10 shared that they changed the time-related examples inside the hook as LLMs sometimes lacked updated information. Hence, they replaced the LLM output with a more recent example.

For R3 (sparking curiosity), P1 and P4 shared that they know what makes a tweet go viral and get clicks from their past Twitter experiences: using exaggeration, shock factors, and potentially misleading information. Then, P4 prepended *"Apparently we're gonna lose $10.5 trillion to criminals over the internet by 2025. Isn't that horrendous?"* to the LLM-generated hook on cybercrime. They believed the addition of surprising data would attract readers more.

5. Users Edit LLM Hooks for Personal Style Users also edited the LLM-generated hooks to make them fit more according to their writing styles and favorite examples so they felt more connected and related to their hooks. For example, P10 shared that they wanted to use the exact syntax they used daily in this hook. So they changed a lot of word-level choices like from "*Do you know what*" to "*Have you ever heard.*" P10 also shared that they intentionally deleted words like "exactly" and split the two questions which were originally in one sentence into two separate short ones. From this, P10 shared that it made them feel that the hook sounded like themselves or their friends by referring to their usual language choices. Also, P1 and P10 edited all of the LLM-generated hooks when they reached Step 6, even though they stated they were already highly satisfied with them. They still expressed wanting to embed more of their styles inside the hook. P4 suggested that making these changes helped maintain their own voice, and P6 specifically added several hashtags and emojis as they liked them. According to P8, engaging in the final editing of the hooks helped them feel a greater sense of agency and ownership over them. This was because they perceived the final product as being more original after undergoing the editing process. P8 specifically mentioned while editing, they shifted from the role of "creator" to the "first reader" of the hooks. By doing so, they gained a more objective and distant view of their writings.

Discussion and Future Works

In this paper, we demonstrate that LLMs can help contextualize technical information into relatable and engaging hooks. We scaffold the complex Tweetorial hook writing process by prompting LLMs for everyday examples, common experiences, and specific anecdotes. This scaffolding approach (MacNeil et al. 2021) helps STEM experts effectively communicate science to non-technical audiences. In the future, it is possible that similar tools could be built for other groups of experts, such as helping journalists reach younger audiences, helping medical professionals explain procedures to patients, or helping public service organizations spread messages to under-served communities.

However, LLMs are far from perfect and user interaction is essential to producing successful hooks. LLMs sometimes provide inaccurate examples for a topic and sometimes suggest experiences that a non-technical audience would not relate to, such as building a website or buying something on the dark web. Ultimately, the expert must decide whether the suggestions are correct and appropriate, and they cannot just "trust the machine." Experts have the ability to judge whether the examples of the technology are correct (such as verifying that Spotify Wrapped does indeed use an API), but they might not understand non-technical audiences well enough to evaluate whether the suggested experiences resonate with them (such as being aware of a lawsuit between Oracle and Google). If an expert is unsure whether something would resonate with the public, they should ask members of their audience. One feature that could be built into such as system is to get human judgments from an online marketplace to provide audience feedback on demand.

Although LLMs have a wealth of information, they do contain biases and not all viewpoints are equally represented. For explaining science to the general public, the biases in the current LLMs like GPT-3 and GPT-4 are probably not problematic. However, if the intended audience were a more specific demographic, LLMs might not suggest examples and experiences that resonate with them. People of different ages, cultural backgrounds, education levels, language abilities, and geographic locations communicate very differently. For example, an experience about using a laptop might not resonate with a low-income student who cannot afford a laptop and does all of their computing from a phone. Currently, LLMs mostly echo dominant perspectives, but it could be powerful to train LLMs to elevate the voices of non-dominant groups as a means to bridge the gap, better support the communities, and promote inclusivity.

In the study, users stated that it was important to them that their final hooks reflected their own personal style and creativity. This is in line with previous work on the social dynamics of AI co-creative systems (Gero, Long, and Chilton 2023) which has shown that when working with LLMs, writers care deeply about preserving their *intent* and the *authenticity* of their writing. To further enable this, some users suggested future versions of the system where writers can feed their hooks back to their system to "keep" their style for future generations, or add a "temperature" parameter to control the specificity of contextualized examples. These features can provide a range of agency when co-creating hooks with LLMs, thus aligning with the future vision of designing more user-focused interactive creativity support tools. These designs can empower users in their content creation process by fostering a sense of ownership and creative expression.

Conclusion

This paper explores integrating generative AI into the hook writing process for Tweetorials, a science communication method that motivates science through relatable examples and experiences. Our prompting engineering study suggests that including examples of good hooks in the prompt helped LLMs generate better hooks, but there is still a need for humans in the loop. To help experts write hooks, we built an LLM-based workflow that scaffolds the process: given a topic, the system suggests everyday examples of the topic, and the user can accept a machine suggestion, edit a machine suggestion, request more suggestions, or write their own. Based on the everyday example selected by the user, the system suggests common experiences. The user can again accept, edit, regenerate, or write their own. Based on the common experience selected, the system suggests a personal anecdote and can make the anecdote more specific while the user may edit these as well. Finally, the system produces an example hook that users can accept as is, or reference when finalizing their hook. Our user study shows this scaffolding greatly reduces the cognitive load of writing hooks. Also, as the outputs are editable at every stage, the hooks still convey the writer's authentic style, voice, and experiences.

Author Contributions

TL finalized the prompt engineering works, built the system, led the annotation and user study, analyzed the results, and wrote the paper. DZ led the early prompt engineering and task understanding and contributed to the system and data collection. GL, BT, SM, and KS assisted with the early task understanding, data collection, and analysis. SW, KG, and LC provided overall guidance on the project, helped shape the two studies, and contributed to the writing.

Acknowledgments

This work has been supported by NSF-IIS-2129020 and NSF-EAR-2121649.

Appendix

Due to the page limits, a high-resolution source appendix is linked here: `https://tinyurl.com/tweetappendix`

Score	1	2	3	4	5
Jargon-Free: Does the hook avoid jargon or unexplained terminology so the general audience can understand it easily?	**Very hard to understand; includes unexplained terminology and most readers cannot understand** **Example:** Website takes forever to load? That's because domain name servers (DNS) are down or not responding. DNS servers are used to translate web addresses (like Google.com) into IP addresses that can be read and understood by computers. 1/ **Reason:** *"'IP addresses"*, "*translate web addresses"*, and "*domain name servers"* are all words most everyday people don't know. These words makes the meaning of this hook unclear.	**Somewhat hard to understand; includes unexplained terminology but some readers can get the gist** **Example:** Slow your internet has been while streaming your favorite show? What if Random Internet Congestion was actually caused by someone else? Sure might make you mad. So what's a DDoS attack, and how can you protect yourself? Read on to find out! 1/ **Reason: "***Random Internet Congestion"* and "*DDoS attack"* are words most everyday people don't know. The reader must dig deep into context clues to understand the hook.	**Somewhat easy for most readers to understand; includes unexplained terminology but most readers can understand** **Example:** With 2 Factor Authentication, you not only have to know the password to get into an account, you also need a second piece of information that only you should know. 1/ **Reason:** This mentions an unfamiliar term first, but follows up with minimal explanation. "*piece of information"* is also unclear.	**Somewhat easy for all readers to understand; includes unexplained terminology but all readers can understand** **Example:** Have you ever had to submit your private medical information to your insurance? How do they keep it secure, yet allow more people to access it? Deidentification is a process that can make sure our information remains private – but how does it work? 1/ **Reason: "***Deidentifcation"* is a word most people don't know, but general context is provided so the reader can guess the meaning. However, a precise definition isn't given.	**Very easy for all readers to understand; does not have any unexplained terminology** **Example:** Have you ever forgotten your password? Gmail sends you a text message with a long number in it, and helps you stay safe! Let's read #2FactorAuthentication. **Reason: "***password"*, "*Gmail"*, and "*text message"* are all terms people are familiar with. Meaning of hook is clear. "#2FA" is jargon but it only shows at the end and serves as an opening for the following tweets.
Specific & Relatable Example(s): Does the hook include a specific and relatable example(s) about the topic?	**Has NO example** **Example:** What are Domain Name Servers & why do they matter? Here's everything you need to know: **Reason:** There is no example.	**Provides an example that is extremely not specific or relatable** **Example**: Data security is important! #Deidentification strips personal info from data while preserving its structure and insights. Have you ever wondered how companies can analyze data without compromising your privacy? #DataPrivacy **Reason**: There is one example, "*stripping personal info from data"*, but it is not specific or relatable.	**Provides an example that is both somewhat specific & relatable** **Example 1:** Have you ever upload a file to the internet and wondered how it gets from your computer to the site's server? Cloud Computing! **Example 2:** I recently witnessed the power of exploiting vulnerabilities first-hand when my team of hackers used a buffer overflow attack to gain access to an exposed Windows 2003 server. Here's an exploration of the world of hacking: 1/ **Reason:** For example 1, "*upload a file*" is a relatable example but it is not specific. For example 2, "*buffer overflow attack of a Windows 2003 server*" is specific, but it's not relatable.	**Provides an example that is specific & relatable for many readers, but not all** **Example:** Have you ever been tempted to try to 'hack' something? Recently, I had a friend try to access my math teacher's laptop in an attempt to improve his grade. But in the end, was it really worth it? Regardless of the outcome, I'm still curious as to why some people would risk so much to hack! Let's explore together. #hacking 1/ **Reason:** The example is specific and relatable. But only some readers may connect with it, not all.	**Provides an example that is specific & relatable for almost all readers** **Example:** I once torrented Last Week Tonight -- then my landlord got a complaint from Comcast! Wth? My friends never got caught. Ugh. So here are things I wished I had known about how to be sneaky on the internet: 1/ **Reason:** The example is intriguing and this experience is common and relatable to almost everyone. Also, "*Last Week Tonight"*, "*landlord"*, "*Comcast", and "my friends*" are detailed enough to make people feel vivid.
Sparks Curiosity: Does the hook give readers a specific and driving reason to keep reading to satisfy their curiosity?	**The tweet does not generate curiosity for readers, OR it has a good question, but it is answered in detail.** **Example:** With 2 Factor Authentication, you not only have to know the password to get into an account, you also need a second piece of information that only you should know. 1/ **Reason:** This is a statement; there is no question. It also directly explains the term and doesn't prompt further questioning.	**The tweet may generate mild curiousity for a small group of people.** **Example:** When someone floods your website with too much traffic that it crashes, that's what's known as a #DDoSAttack. Have you ever had to deal with a similar scenario? What did you do? #CyberSecurity #Websecurity **Reason:** Only a select group may be curious about crashing a website through too much traffic. Hook doesn't present a specific or urgent question.	**The tweet generates some curiosity for readers, OR it has a good question but provides too much of an answer** **Example 1:** Have you ever uploaded a file to the internet and wondered how it gets from your computer to the site's server? Cloud Computing is the answer! **Example 2:** My son relies on his Alexa to help with his math homework every single night. While I am concerned about his learning, I am interested in how it works. **Reason:** These hooks have questions and examples to intrigue people, but they are not specific enough to make people feel very curious.	**The tweet may generate curiousity for many readers, but not all** **Example:** I just bought some things online, but how do I know the website I'm using is safe? Without HTTPS, anyone can intercept the data I send out...but how does HTTPS keep me protected when I'm online shopping? Here's what I've learned so far about online security: 1/ **Reason:** The hook identifies general questions that many people might have, but does not have a specific question.	**The tweet instills curiosity and makes you want to read more.** **Example 1:** I once torrented Last Week Tonight -- then my landlord got a complaint from Comcast! Wth? My friends never got caught. Ugh. So here are things I wished I had known about how to be sneaky on the internet: **Example 2:** My son relies on his Alexa to help with his math homework every single night. While I am concerned about his learning, I am interested in how Alexa understands what he is saying? Is it the same way that humans understand language? **Reason:** These hooks provide the reader with a specific question that grabs their attention and makes them want to continue reading.

Figure 6: A five-point scale rubric for annotators in the prompt engineering study

Ransomware	Database	Browser Hijacking	Machine Learning	**API***	Patch	White Hat	**Programming Language***	Trojan	Ad Blocking
Front End*	Peer-To-Peer	**Net Neutrality***	Internet Service Provider	Tor	Black Hat	BitTorrent	Secure Socket Layer	**Cybercrime***	Troll
Back End	Browser	Breach Notification	Access Control List	Dark Web	OAuth	Security Key	**Autocomplete***	Backdoor	Spoofing

Table 3: 30 topic inputs for the prompt engineering study (*asterisked are the six topics randomly chosen for the user study)

Strategy	Lowest-scoring hook	Lowest-scoring hook with same topic as highest-scoring hook	Highest-scoring hook
PS1	**Application Programming Interface (API)** A #programming interface that lets you calculate prices, map distances, and get stock quotes automatically? What else can an #API do and how can it help me? **Average Score: 1.89**	**Browser Hijacking** A hacker could take control of your online activity and webpages without you realizing it! Ever heard of browser hijacking? Have you taken the necessary steps to protect yourself? #browserhijacking #cybersecurity #hackerrisk **Average Score: 2.89**	**Browser Hijacking** Beware of browser hijacking! Ever gone to a website you didn't type in and been redirected somewhere else? That's browser hijacking! It can be used for malicious purposes - so how can you stay safe? #browserhijacking #securty **Average Score: 4.00**
PS2	**The Onion Router (TOR)** Ever hear about the deep web and the black markets? Well, TOR is the gateway to them! What's required to risk taking a journey into the unknown? A thread on TOR and its implications: **Average Score: 1.67**	**Internet Service Provider** Have you ever wondered why you get more slow internet in your bedroom than at the dining table? What is it that your internet service provider does differently to affect your internet connection? Let's look into it: 1/ **Average Score: 3.44**	**Internet Service Provider** Have you ever had an internet outage at the exact moment when you need to send an assignment or video chat with friends? Who's responsible for this? How are we connected to the internet? I'm about to explore what happens behind the scenes of the internet. Let's start by understanding the main characters involved - the Internet Service Providers. Ready? **Average Score: 4.55**
PS3	**Back End** Have you ever wanted to keep your data secure and away from prying eyes? You might not realize it, but controlling access to your accounts is just as important as data privacy. But how can you do this? My recent experience with Amazon Web Services' Identity and Access Management feature has shown me the power of properly managing back end access. Let's discover more! 1/ **Average Score: 2.11**	**Internet Service Provider** Have you ever wondered why an Internet Service Provider (ISP) could make promises that never turn out to be true? I recently learned the hard-way when my streaming experience got slower even though I had chosen a top-tier plan from a big-name provider. Here's an exploration of how ISP's work and why you should take their promises with a grain of salt. #ISP #Technology #Internet #Promises 1/ **Average Score: 3.55**	**Internet Service Provider** I used to think living in the suburbs would mean better internet connection. But my Verizon Fios service was so unreliable and slow that I felt like I was back in the dark ages! What can we do to better understand the nature of internet service providers, and how can they provide truly reliable service? Here's the story: **Average Score: 4.78**

Table 4: Collection of good and bad hooks from the prompt engineering study

References

Aldous, K. K.; An, J.; and Jansen, B. J. 2019. The Challenges of Creating Engaging Content: Results from a Focus Group Study of a Popular News Media Organization. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 1–6. Glasgow, Scotland, UK: Association for Computing Machinery.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. Virtual Event Canada: Association for Computing Machinery.

Breu, A. C. 2019. Why Is a Cow? Curiosity, Tweetorials, and the Return to Why. New England Journal of Medicine 381(12):1097–1098.

Breu, A. C. 2020. From Tweetstorm to Tweetorials: Threaded Tweets as a Tool for Medical Education and Knowledge Dissemination. Seminars in Nephrology 40(3):273–278.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs]. arXiv: 2005.14165.

Brüggemann, M.; Lörcher, I.; and Walter, S. 2020. Post-normal science communication: exploring the blurring boundaries of science and journalism. Journal of Science Communication 19(03):A02.

Bukhtiyarov, A., and Gusev, I. 2020. Advances of transformer-based models for news headline generation.

Bullock, O. M.; Colón Amill, D.; Shulman, H. C.; and Dixon, G. N. 2019. Jargon as a barrier to effective science communication: Evidence from metacognition. Public Understanding of Science 28(7):845–853.

Calderwood, A.; Wardrip-Fruin, N.; and Mateas, M. 2022. Spinning coherent interactive fiction through foundation model prompts. In ICCC, 44–53. Association for Computational Creativity (ACC).

Chung, J. J. Y.; Kim, W.; Yoo, K. M.; Lee, H.; Adar, E.; and Chang, M. 2022. Talebrush: Sketching stories with generative pretrained language models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22. New York, NY, USA: Association for Computing Machinery.

Crist, M. 2019. "I've been watched the landscape of my childhood burn with an aching heart, and wondering how much climate change is to blame. Turns out human activity is a major driver of California's wildfires, but not just in the ways you might imagine.". `https://twitter.com/meehancrist/status/1197527975379505152`. [Online; accessed June-2022].

Emig, J. 1977. Writing as a Mode of Learning. College Composition and Communication 28(2):122–128. Publisher: National Council of Teachers of English.

Flower, L., and Hayes, J. R. 1981. A Cognitive Process Theory of Writing. College Composition and Communication 32(4):365.

Gero, K. I.; Liu, V.; Huang, S.; Lee, J.; and Chilton, L. B. 2021. What makes tweetorials tick: How experts communicate complex topics on twitter. Proc. ACM Hum.-Comput. Interact. 5(CSCW2).

Gero, K. I.; Liu, V.; and Chilton, L. 2022. Sparks: Inspiration for science writing using language models. In Designing Interactive Systems Conference, DIS '22, 1002–1019. New York, NY, USA: Association for Computing Machinery.

Gero, K. I.; Long, T.; and Chilton, L. B. 2023. Social dynamics of AI support in creative writing. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. New York, NY, USA: Association for Computing Machinery.

Hart, S. G., and Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. Human mental workload 1(3):139–183.

Hayes, J. R. 1996. A new framework for understanding cognition and affect in writing. In The Science of Writing: Theories, Methods, Individual Differences, and Applications. Lawrence Erbaum Associates.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. arXiv:1904.09751 [cs].

Howell, E. L.; Nepper, J.; Brossard, D.; Xenos, M. A.; and Scheufele, D. A. 2019. Engagement present and future: Graduate student and faculty perceptions of social media and the role of the public in science engagement. PLOS ONE 14(5):e0216274.

Ippolito, D.; Kriz, R.; Kustikova, M.; Sedoc, J.; and Callison-Burch, C. 2019. Comparison of Diverse Decoding Methods from Conditional Language Models. arXiv:1906.06362 [cs]. arXiv: 1906.06362.

Jigsaw. 2017. Sideways dictionary. `https://sidewaysdictionary.com/#/`.

Jin, D.; Jin, Z.; Zhou, J. T.; Orii, L.; and Szolovits, P. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5082–5093. Online: Association for Computational Linguistics.

MacNeil, S.; Ding, Z.; Quan, K.; Parashos, T. j.; Sun, Y.; and Dow, S. P. 2021. Framing creative work: Helping novices frame better problems through interactive scaffolding. In Creativity and Cognition. New York, NY, USA: Association for Computing Machinery.

McClain, C., and Neeley, L. 2014. A critical evaluation of science outreach via social media: Its role and impact on scientists. F1000Research.

Petridis, S.; Diakopoulos, N.; Crowston, K.; Hansen, M.; Henderson, K.; Jastrzebski, S.; Nickerson, J. V.; and Chilton, L. B. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. New York, NY, USA: Association for Computing Machinery.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. OpenAI blog 1(8):9.

Shearer, E., and Matsa, K. E. 2018. News Use Across Social Media Platforms 2018. Pew Research Center.

Singh, N.; Bernal, G.; Savchenko, D.; and Glassman, E. L. 2022. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. ACM Trans. Comput.-Hum. Interact. Just Accepted.

Wang, S.; Petridis, S.; Kwon, T.; Ma, X.; and Chilton, L. B. 2023. Popblends: Strategies for conceptual blending with large language models. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23. New York, NY, USA: Association for Computing Machinery.

Wu, T.; Terry, M.; and Cai, C. J. 2022. AI Chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22. New York, NY, USA: Association for Computing Machinery.

Xu, P.; Wu, C.-S.; Madotto, A.; and Fung, P. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning.

Yeo, S. K. 2015. Public engagement with and communication of science in a web-2.0 media environment. Washington, DC: The American Association for the Advancement of Science (AAAS).

2. Co-creativity

The Emperor's New Co-Author

Dan Ventura*
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
ventura@cs.byu.edu

Abstract

We examine the potential of AI language models as co-creative partners and discuss the limitations of current models like ChatGPT. While there is excitement around these models, they currently lack understanding, cannot produce high-quality writing, and cannot engage in dialogue. Understanding language goes beyond language modeling, and the symbol grounding problem highlights why language models cannot understand language. This lack of understanding leads to poor writing, and while AI-generated content is easy to detect, sorting through the abundance of low-quality content is a growing concern. Further, co-creative partnerships require dialogue, and AI models are not currently capable of this, because they do not possess, among other things, the ability to understand (a partner's) processes and goals. Our conclusion is that AI language models are incapable of acting as co-creators and relying on them as such may actually hinder human creativity rather than complement or augment it.

Introduction

In the past few years, artificial Intelligence (AI) has experienced numerous breakthroughs in research and significant advancements in practical applications. The proliferation of powerful computing resources, vast amounts of data, and advanced algorithms has facilitated the widespread adoption of AI across different industries, leading to increased productivity, efficiency, and innovation (Unknown1 2021; Unknown2 2023; Unknown4 2019). In the last decade, venture capital investment in AI startups has soared, with billions of dollars being pumped into the sector (Unknown3 2022). Major tech companies such as Google, Microsoft, and Amazon have also been investing heavily in AI, acquiring startups and expanding their AI capabilities through internal research and development efforts.

AI is being used to improve healthcare, finance, transportation, education, and many other sectors. For instance, AI-powered algorithms are being used to diagnose medical conditions more accurately and quickly (Unknown5 2019), and machine learning algorithms are being used to analyze financial data to detect fraudulent activities (Unknown6 2022). In transportation, self-driving cars are becoming more prevalent, and in education, AI is being used to personalize learning experiences for students (Unknown7 2018; Zeng et al. 2021).

The increasing availability of data and computing resources is also accelerating the development and adoption of AI. With the proliferation of sensors and the Internet of Things (IoT), vast amounts of data are being generated, providing fertile ground for machine learning algorithms to learn from (Unknown8 2022). Furthermore, the increasing availability of cloud computing services and powerful processors has made it easier and more cost-effective to train and deploy AI models.

ChatGPT is a state-of-the-art language model developed by OpenAI, which has recently captured the spotlight due to its impressive performance and capabilities (Radford et al. 2019; Brown et al. 2020). With a staggering 1.6 billion parameters, ChatGPT is currently one of the largest and most powerful language models available. On the SuperGLUE benchmark (Wang et al. 2019), ChatGPT achieved a score of 89.8, which is currently the highest score achieved by any language model on this benchmark. Additionally, ChatGPT has been shown to perform well in other language tasks, such as machine translation, summarization, and question-answering. ChatGPT can be fine-tuned to perform specific language tasks, making it highly adaptable to different use cases. It has been used in a wide range of applications, from chatbots and customer service systems to educational tools and language translation services. Furthermore, the availability of ChatGPT's pre-trained model and open-source code has made it accessible to a wider audience of developers and researchers, allowing for further advancements and applications of this technology.

The development of advanced AI language models is seen by many as a significant milestone in the field of artificial intelligence. These models are touted as a major step forward in the ability of machines to understand and generate human language, which has long been considered one of the most challenging tasks in AI, and they have demonstrated remarkable performance in a wide range of natural language processing tasks, including language translation, text summarization, and even creative writing (Hossain, Shrestha, and

*This is a paper-within-a-paper. The main paper was written with heavy use of ChatGPT. The meta-paper/analysis was written solely by me, as I've traditionally done. See Appendix for details.

Yamada 2020; Li and Li 2020; Sun, Cai, and Ren 2020).

The possibility for AI language models to enhance creativity and productivity is tantalizing. With their ability to generate high-quality text, these models could provide valuable support to writers, journalists, and other content creators, allowing them to quickly and efficiently generate ideas and drafts. They could also help researchers and scientists to analyze large amounts of text data, leading to new insights and discoveries (Khosla 2020).

Perhaps even more enticing is the idea of these models acting as a co-creative partner with humans (Dodge et al. 2021; Mubin, Bartneck, and Feijs 2020). Instead of simply generating text, could these models actively collaborate with humans in the creative process, generating ideas, providing feedback, and enhancing the overall quality of the final output? This idea of co-creation is particularly exciting in creative industries such as writing, music, and art, where collaboration between individuals with different perspectives and skill sets can often lead to innovative and inspiring results (Karamcheti 2021; Nirenburg 2020). With the help of AI language models, could this collaborative process become even more powerful and efficient, allowing creators to explore new ideas and push the boundaries of their respective fields?

In addition to creative industries, the idea of co-creation has potential applications in other areas such as education, scientific research, and healthcare. For example, AI language models could help students and teachers to collaborate on writing assignments, providing suggestions for improvements and generating new ideas.

However, it is important to note that the idea of co-creation between humans and AI language models is still in its early stages and faces many challenges, perhaps the most significant being the ability of the AI model to understand and adapt to the unique preferences and creative styles of its human partner. Indeed, we take the position that *current AI language models are fundamentally incapable of acting as a collaborator* for at least three critical reasons:

1. AI language models lack the capacity for genuine understanding. They may be able to generate language patterns based on statistical analysis of training data, but they lack the contextual and emotional intelligence required for true understanding. This makes it difficult for them to contribute meaningfully to collaborative projects, as they cannot fully comprehend the goals, perspectives, and experiences of human collaborators.

2. AI language models often struggle with producing writing that meets the standards of quality expected in collaborative projects. While they may be able to generate language that is grammatically correct and semantically coherent, their writing is often lacking in creativity, style, and voice. This can make it difficult for them to contribute meaningfully to collaborative projects, as they cannot produce writing that matches the quality and style of human collaborators.

3. AI language models are typically unable to engage in dialogic exchanges that are central to true collaboration. While they may be able to generate language in response to prompts, they lack the capacity for genuine dialogue. This means that they cannot engage in the back-and-forth exchanges of ideas, feedback, and revision that are central to collaborative projects.

Understanding vs. modelling

Language modeling is the task of assigning probabilities to sequences of words in a language. A language model is a mathematical model that captures the distribution of word sequences in a language. Given a sequence of words, $w_1, w_2, ..., w_n$, a language model calculates the probability of this sequence, $P(w_1, w_2, ..., w_n)$, as the product of the probabilities of each word given the context of the preceding words

$$P(w_1, w_2, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \\ \ldots P(w_n|w_1, w_2, ..., w_{n-1})$$

where $P(w_i|w_1, w_2, ..., w_{i-1})$ is the probability of word w_i given the context of the preceding words $w_1, w_2, ..., w_{i-1}$. The language model is trained on a large corpus of text by estimating the probabilities of word sequences based on their frequency of occurrence in the training data. The goal of language modeling is to accurately predict the likelihood of a sequence of words.

Whereas language modeling is the task of assigning probabilities to sequences of words in a language, understanding refers to the ability to comprehend and interpret meaning from such sequences. Understanding involves processing linguistic input and making sense of it based on a range of contextual, pragmatic, and semantic factors. It involves higher-level cognitive processes such as inference, reasoning, and background knowledge, which go beyond the statistical patterns of language modeling. Understanding involves not only recognizing individual words and their syntactic structure but also grasping the intended meaning, discourse structure, and communicative goals of the speaker or writer.

Recent advances in AI have come almost exclusively through huge increases in available training data and computational power. However, increasing the amount of data and computational resources used for language modeling does not lead to an increase in language understanding— it just means that the language model can better capture the statistical patterns of the language and produce more accurate predictions of word sequences. To achieve language understanding, additional techniques such as semantic analysis, knowledge representation, and reasoning are needed. These techniques often require explicit modeling of linguistic and world knowledge, as well as more advanced reasoning and inference mechanisms.

The symbol grounding problem is a philosophical and computational problem in artificial intelligence and cognitive science that arises from the fact that symbols or words in natural language lack inherent meaning (Harnad 1990) . The problem can be stated as follows:

> Given a set of symbols or words, how can a machine or a cognitive agent associate them with their corresponding meanings in the physical world? In other words,

how can symbols be grounded or anchored to the external world, such that they can be used to represent and reason about real-world entities and events?

It is particularly relevant to natural language processing and understanding, as language relies heavily on symbols or words to represent concepts and convey meaning. However, the meanings of these symbols are not inherent in the symbols themselves, but are rather derived from their use in the context of real-world situations and experiences, requiring a machine or a cognitive agent to be able to perceive and interact with the physical world in a meaningful way, and to learn the correspondences between symbols and their associated meanings through experience and interaction.

Because language models are based on statistical models that learn to predict the probability distribution of words in a given context based on their previous occurrences in large corpora of text, they do not have any understanding or knowledge of the real-world entities or events that the words represent, but rather rely on the co-occurrence patterns of words in the data to make predictions. This is obviously problematic when it comes to understanding natural language, because language is not just a collection of words, but a means of representing and communicating about real-world entities and events. Words and symbols in natural language are grounded or anchored to the external world through a process of association and learning that involves perception, action, and experience.

The symbol grounding problem is relevant here because language models lack this grounding and do not have any direct connection to the real-world entities and events that the words represent. Language models are not able to perceive or interact with the external world, and therefore cannot derive the meanings of the words from their context in the world (Barsalou 1999; Glenberg and Robertson 2000).

However, even if such language models could somehow address the problem, symbol grounding alone is not enough for achieving true language understanding because it only addresses the individual aspect of symbol grounding. For language to be truly understood, there must be a common/shared grounding between the speakers or agents communicating. This means that the symbols or words used in language must be grounded in a shared external reality, so that they can be understood and communicated between different individuals or agents. A shared grounding in external reality allows for a common understanding of the meanings of words and symbols, and for the ability to refer to the same entities and events. Without a shared grounding, language becomes a collection of individual associations between symbols and personal experiences, which cannot be effectively communicated or understood by others.

Skill, or lack thereof?

Language understanding is a critical component of effective writing. Poor language understanding can result in inadequate writing, making it challenging for the writer to convey their message effectively. A writer who lacks language understanding may make mistakes in grammar, leading to sentences that are difficult to understand, contain incomplete thoughts, or are confusing to the reader. A writer who lacks language understanding may struggle with word choice. Vocabulary is an essential component of effective writing, and a writer must have a strong grasp of the meanings of words to use them correctly. Without language understanding, a writer may struggle with organizing their thoughts, leading to writing that is disjointed or lacks coherence. A writer who lacks language understanding may struggle with choosing an appropriate style and tone for their writing, leading to writing that is inappropriate or ineffective.

Proper grammar, a good vocabulary, and correct punctuation and spelling are necessary (except when they are not) but not sufficient (except when they are) to guarantee good writing. Good writing (almost always) requires more than just technical correctness. Technically correct writing can be stiff, formulaic, or lacking in personality. Good writing should be engaging and interesting to read, which often requires the use of creative and unconventional sentence structures, word choices, and rhetorical techniques. Focusing too heavily on technical correctness can lead to writing that is overly cautious and lacking in voice or personality. It can also overlook the importance of context and audience. Good writing is written with a specific audience and purpose in mind, and the language, tone, and style of the writing should reflect those considerations. Simply using proper grammar and vocabulary does not guarantee that the writing will be effective in achieving its intended purpose or connecting with its intended audience. Good writing requires not only technical correctness but also a clear and coherent message that engages the reader and communicates ideas effectively. Focusing too heavily on technical correctness can lead to writing that is dry and formulaic, and lacks the depth and substance required to engage and inform the reader.

AI/language models, while capable of generating coherent and grammatically correct sentences, often produce writing that is banal and shallow because they lack true understanding of the meaning and nuance of language. Analyzing large amounts of data and learning patterns to predict and generate new text does not involve true comprehension of the meaning of the language being generated. Language models lack the ability to understand context and the complexities of human experience. Writing that is truly engaging and thought-provoking often requires an understanding of the underlying meaning and nuance of language, as well as an ability to interpret and respond to the specific context in which the writing is being produced. AI/language models, lacking understanding, cannot effectively respond to context in the same way that human writers can. Language models often rely on formulaic language and patterns, leading to repetitive and predictable writing. While this may produce text that is grammatically correct, it can also lead to writing that is bland and lacking in creativity. Good writing requires not only technical correctness but also originality and an ability to engage the reader with fresh ideas and perspectives. AI/language models, lacking understanding and creativity, may struggle to produce writing that captures the reader's attention. Language models lack the human perspective and insight that is necessary for engaging writing. Writing that is thought-provoking often draws on personal experience and insight, and requires a deep understanding of the human experience. While AI/language models can produce text that is superficially similar to human writing, they lack the depth and insight that comes from true human experience and understanding.

Despite this, there is growing concern about the potential for people to cheat in various contexts by using language models or other AI tools to generate text that appears to be their own work. The widespread availability of language models and other AI tools that can produce coherent and grammatically correct text has made it easier than ever for individuals to produce written content quickly and easily, potentially giving them an unfair advantage in academic or professional contexts:

- Students may use language models to generate essays or other written assignments, presenting them as their own work without fully understanding the material or demonstrating their own critical thinking skills. This can lead to a devaluation of academic standards and undermine the integrity of educational institutions.
- Professionals may use language models to produce reports or other documents, presenting them as their own work without truly understanding the material or conducting the necessary research. This can lead to errors and inaccuracies in important documents, potentially causing harm to individuals or organizations that rely on them.
- Language models may be used to generate fake news or other misleading information, further eroding trust in information sources and undermining public discourse.

While some may believe that it is difficult to detect AI-generated content, in reality, for the reasons give above, it is usually relatively simple: AI-generated content often lacks the nuance and depth of human-generated content; tends to be formulaic and repetitive; lacks originality and creativity; and often exhibits patterns that can be detected through machine learning algorithms or other analysis techniques.

However, while the detection of AI-generated content may not be particularly challenging, the real danger lies in the sheer volume of poor writing that is produced as a result of the widespread availability of language models and other AI tools. The ease and speed with which these tools can produce written content has led to a glut of low-quality writing, much of which is difficult to sift through and evaluate. This can be particularly problematic in contexts such as online publishing, where there is a high demand for content and a need to produce it quickly and efficiently. As a result, much of the content that is produced is of low quality, lacking originality and insight and often containing errors and inaccuracies. The cost of sorting through this glut of poor writing can be significant, both in terms of time and resources. This can place a burden on those tasked with evaluating or curating content, and can also lead to a devaluation of high-quality writing and a lowering of standards.

Dialogic disability

A co-creative partner requires some kind of dialogic ability. Co-creation involves the joint creation of something, whether it be a product, service, or experience, and requires a collaborative effort between two or more parties. For this collaboration to be successful, it is necessary for all parties to have the ability to engage in a dialogue or conversation.

In a co-creative partnership, both parties bring their own unique knowledge, skills, and perspectives to the table. The ability to engage in a dialogue allows these parties to share their ideas and insights with one another, build on each other's contributions, and work together to create something that is greater than the sum of its parts.

Without a dialogic ability, a co-creative partnership can become one-sided or unproductive. If one party dominates the conversation or is unwilling to listen to the ideas and perspectives of the other party, the collaboration can quickly become imbalanced and unproductive. In addition, the ability to engage in a dialogue is important for building trust and fostering a sense of shared ownership in the co-creative process. When both parties are able to contribute and participate equally in the collaboration, they are more likely to feel invested in the outcome and committed to its success.

People are unlikely to grant partner status to an entity that doesn't understand their process or goals. Partnerships involve a shared commitment to a common goal, and require a high degree of collaboration, communication, and mutual understanding— a lack of understanding creates a barrier to effective communication, collaboration, and mutual support, which are essential components of a successful partnership. Without this shared understanding, it is difficult to build trust and establish a sense of shared ownership in the partnership. This can lead to a breakdown in communication, misunderstandings, and a lack of coordination, which can ultimately result in the failure of the partnership.

While it is possible for an AI language model to assist with certain aspects of the writing process, such as grammar and sentence structure, it cannot take on the subtask of writing the next section, drawing a figure, or formalizing an algorithm. This is because the AI lacks the contextual understanding and knowledge necessary to make informed decisions about the structure and content of the work.

For a real co-author, taking on the subtask of writing the next section, drawing a figure, or formalizing an algorithm requires an understanding of the topic and the goals of the work. This understanding allows the co-author to make informed decisions about the content, structure, and presentation of the work. Additionally, the co-author can engage in a dialogue with the other co-authors to ensure that their contributions are aligned with the overall goals of the project. An AI language model lacks the contextual understanding and knowledge necessary to make such informed decisions. While it may be able to generate text based on a given prompt, it cannot make decisions about the next section or topic of the work without a deeper understanding of the project as a whole. Similarly, an AI language model cannot create figures or formalize algorithms without an understanding of the underlying concepts and their relevance to the work.

Language models may be considered the calculators of writing. Just as calculators are tools that help us perform complex mathematical calculations quickly and accurately, language models are tools that can generate written content with remarkable speed and precision. They can assist with various writing tasks, such as grammar and syntax correction, sentence structure improvement, and generating entire pieces of text based on a prompt. However, just as calculators are limited in their ability to solve complex mathematical problems that require creative thinking and problem-solving skills, language models have their limitations as well. While they can generate text quickly and accurately based on a given prompt, they lack contextual understanding and creativity. Language models cannot replicate the nuances of human communication, such as humor, irony, and sarcasm, which are essential elements of effective writing. Further,writing involves a creative and iterative process that requires critical thinking, problem-solving, and the ability to make informed decisions about the content, structure, and tone of the writing.

ChatGPT and similar language models are powerful tools that can assist with various writing tasks, but they cannot be considered co-authors any more than a calculator can be considered a co-inventor of a proof in mathematics. Another analogy for language models might be a theorem prover— . they do not contribute any novel ideas or insights to the proof, but rather assist with the verification process by systematically checking the proof for errors and inconsistencies.

Evolutionary cul-de-sac

The hype surrounding language models powered by artificial intelligence is concerning because it can create unrealistic expectations about the capabilities of these tools. While language models have made significant advances in recent years, they are still limited in their ability to replicate the complexity and nuance of human communication.

One of the key limitations of language models is their lack of contextual understanding. While they can generate text that is grammatically correct and syntactically coherent, they lack the ability to understand the broader context in which the text is being generated. This means that language models can struggle to accurately represent the nuances of human communication, such as tone, humor, and sarcasm, which are critical elements of effective writing.

Another limitation of language models is their inability to replicate the creativity and problem-solving skills that are required for effective writing. While they can generate text quickly and accurately based on a given prompt, they lack the ability to engage in the critical thinking and problem-solving skills that are required for effective writing. This means that language models cannot replicate the creativity and originality that are essential for producing high-quality written content.

Furthermore, language models are only as good as the data that they are trained on. This means that if the data is biased or limited in some way, the language model may reproduce these biases in its output. This can be particularly problematic when it comes to sensitive topics, such as race, gender, and religion, where the language model may unintentionally reproduce harmful stereotypes and biases.

Computational creativity has been hailed as a promising field that can augment human creativity and lead to innovative solutions to complex problems. However, widespread reliance on language models such as ChatGPT may actually have the opposite effect—stifling human creativity rather than augmenting it: language models rely on large datasets of existing text to generate new content, and, as a result, their output tends to be formulaic and lacking in originality; relying on tools such as ChatGPT, humans may become less confident in their own creative abilities and less willing to take risks, leading to a reduction in the diversity and originality of human-generated creative output; language models' failure to understand the emotional or cultural context of the text can lead to insensitive or inappropriate language being generated by the model.

Recently, a well-known science fiction publisher made headlines by announcing that they would no longer accept submissions due to the overwhelming number of low-quality, obviously AI-generated submissions (Heath 2022). This decision highlights a growing concern among publishers and editors about the impact of language models and other AI tools on the publishing industry. The use of language models such as ChatGPT has become increasingly popular among writers and publishers in recent years. These tools promise to make the writing process faster and more efficient by automating tasks such as generating plotlines, dialogue, and even entire stories. However, the downside of this technology is that it can lead to a flood of poorly-written and unoriginal content, as evidenced by the high number of AI-generated submissions received by the science fiction publisher.

Because these tools lack the ability to truly understand the nuances of language and storytelling, they lack the creativity and originality that comes from a human understanding of language and narrative structure, and, as a result, their output can feel stilted, formulaic, and unengaging. Unfortunately, the ease of use of these tools has led to a proliferation of low-effort writing, with some individuals using them to create large volumes of content quickly without paying attention to quality or originality.

The field of artificial intelligence has seen several cycles of hype and disillusionment over the years, with periods of intense excitement and investment followed by periods of disappointment and decline. While recent advances in machine learning and natural language processing have led to remarkable achievements in language modeling and other areas of AI, it is important to maintain a sober perspective on the current state of the field.

It is likely that the current hype surrounding language models and other AI tools has far exceeded reality, and that we are at risk of tipping into the latest AI winter. Despite the impressive capabilities of tools like ChatGPT. The high expectations placed on language models and other AI tools have led to unrealistic demands for their performance and many companies and investors may soon be disappointed with the practical results of their investments, leading to a significant decline in funding and interest in the field.

References

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22(4):577–660.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Dodge, J.; Gao, A.; Schuurmans, D.; and Courville, A. 2021. Evaluating the ability of neural networks to collaborate. https://arxiv.org/pdf/2101.08560.pdf. *This is a fabricated or mis-attributed reference*.

Glenberg, A. M., and Robertson, D. A. 2000. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language* 43(3):379–401.

Harnad, S. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1-3):335–346.

Heath, T. 2022. Publisher closes to submissions due to high volume of low-quality AI-generated manuscripts. https://themedium.com/the-innovation/publisher-closes-to-submissions-due-to-high-volume-of-low-quality-ai-generated-manuscripts-45638c12d1e. *This is a fabricated or mis-attributed reference*.

Hossain, M. S.; Shrestha, A.; and Yamada, Y. 2020. GPT-Infinite: Multi-turn response generation with GPT-2. https://arxiv.org/pdf/2006.06569.pdf. *This is a fabricated or mis-attributed reference*.

Karamcheti, S. 2021. Why AI still can't write like a human. https://www.wired.com/story/why-ai-still-cant-write-like-a-human/. *This is a fabricated or mis-attributed reference*.

Khosla, V. 2020. AI is transforming the future of creative work. https://hbr.org/2020/08/ai-is-transforming-the-future-of-creative-work. *This is a fabricated or mis-attributed reference*.

Li, J., and Li, X. 2020. Multi-user dialogue system with topic-aware transformer. https://arxiv.org/pdf/2012.03448.pdf. *This is a fabricated or mis-attributed reference*.

Mubin, O.; Bartneck, C.; and Feijs, L. M. 2020. Collaboration with artificial intelligence: An experimental study on the effects of context on collaboration. *International Journal of Human-Computer Studies* 142:102473. *This is a fabricated or mis-attributed reference*.

Nirenburg, S. 2020. Conversational AI: From understanding the dialogue to participating in it. https://doi.org/10.1007/978-3-030-41718-3. *This is a fabricated or mis-attributed reference*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Sun, C.; Cai, Z.; and Ren, J. 2020. An empirical study on Chinese poem generation with GPT-2 model. https://arxiv.org/pdf/2010.07229.pdf. *This is a fabricated or mis-attributed reference*.

Unknown1. 2021. The state of artificial intelligence (AI) in 2021. https://www.forbes.com/sites/forbestechcouncil/2021/01/14/the-state-of-artificial-intelligence-ai-in-2021/. *This is a fabricated or mis-attributed reference*.

Unknown2. 2023. Artificial intelligence: The rise of the machines. https://www.mckinsey.com/featured-insights/artificial-intelligence/the-rise-of-the-machines-ai-in-the-modern-world. *This is a fabricated or mis-attributed reference*.

Unknown3. 2022. Global AI market to reach 190.61 billion by 2025. https://www.marketsandmarkets.com/PressReleases/artificial-intelligence.asp. *This is a fabricated or mis-attributed reference*.

Unknown4. 2019. The pervasiveness of ai and its rapid advancements. https://hbr.org/2019/05/the-pervasiveness-of-ai-and-its-rapid-advancements. *This is a fabricated or mis-attributed reference*.

Unknown5. 2019. Artificial intelligence in healthcare: Past, present, and future. https://www.frontiersin.org/articles/10.3389/fpubh.2019.00115/full. *This is a fabricated or mis-attributed reference*.

Unknown6. 2022. The rise of artificial intelligence in finance. https://www.ft.com/content/1caaece4-08d3-11e9-a3aa-118c761d2745. *This is a fabricated or mis-attributed reference*.

Unknown7. 2018. Ai in education: What does the future hold? https://edtechmagazine.com/higher/article/2018/10/ai-education-what-does-future-hold. *This is a fabricated or mis-attributed reference*.

Unknown8. 2022. Artificial intelligence and the internet of things: How the two technologies are evolving together. https://www.zdnet.com/article/artificial-intelligence-and-the-internet-of-things-how-the-two-technologies-are-evolving-together/. *This is a fabricated or mis-attributed reference*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zeng, J.; Wu, Y.; Wu, Z.; and Liu, X. 2021. Application of GPT-3 language model in education field. *International*

Journal of Emerging Technologies in Learning 16(5):150–159. *This is a fabricated or mis-attributed reference*.

Appendix

This exercise was undertaken as a mechanism for elucidating both the feasibility of using current language models as significant aids in academic writing and the longer-term prognosis for (statistical) language models as co-creative agents (for human partners). In what follows, we briefly explain the process used to produce the position paper[1] and then give some informal analysis and discussion of experiences and observations made throughout that process. These are the steps used to produce the main paper:

1. I ideated prompts as a scaffold for a position paper
2. I organized the prompts into appropriate sections in a likely order (section names chosen by me)
3. ChatGPT generated responses to each prompt in order, and I concatenated the responses to form the full paper
4. I asked ChatGPT for supporting references whenever the content of its response suggested they were necessary
5. I presented the complete prompt outline[2] as a prompt to generate the abstract
6. I presented the complete prompt outline as a prompt to generate the title
7. I proof-read and wordsmithed the full text, both for length and quality, as necessary, as I would for the final stage of any paper I write[3]

Prompts

Each unnumbered, bulleted item in the outline below was used as a prompt to generate, on average, 249.88 words of text.[4] All prompts included the prefix ornamentation "Using formal language,"

1. Introduction
   - discuss that AI is really taking off
   - discuss that ChatGPT is the latest model to capture the spotlight
   - discuss that people are very excited about the potential of these new AI (language) models
   - discuss that one tantalizing possibility is having these language models act as a co-creative partner
   - discuss the position that current AI language models are fundamentally incapable of acting as a collaborator because they lack understanding, cannot produce quality writing and cannot be dialogic
2. Understanding vs. modelling
   - explain language modelling
   - explain how understanding is different
   - explain that more data and more compute doesn't translate language modeling into language understanding
   - explain the symbol grounding problem
   - explain why language models don't address the symbol grounding problem and therefore cannot understand language
   - explain why grounding is not enough—it must be a common grounding
3. Skill, or lack thereof?
   - discuss how lack of language understanding leads to poor writing
   - discuss how use of proper grammar, a good vocabulary and correct punctuation and spelling doesn't mean good writing
   - discuss how AI/language models' writing is banal and shallow because it lacks understanding
   - discuss how people are worried about people cheating by using language models
   - discuss how it is actually very simple to detect AI-generated content and how instead the real danger is the cost of sorting through a glut of poor writing
4. Dialogic disability
   - discuss how a co-creative partner requires some kind of dialogic ability
   - argue that people won't grant partner status to an entity that doesn't understand their process or goals
   - talk about how a real co-author could take the subtask of writing the next section or drawing a figure or formalizing an algorithm, but AI can not—not because these things are inherently human but because it has no idea what is the next section, what is the topic is or what to formalize
   - argue that language models are the calculators of writing
   - argue that ChatGPT and its ilk are no more capable of being considered co-authors any more than a calculator should be considered co-inventor of a proof in mathematics; maybe a theorem prover is an even better analogy
5. Evolutionary cul-de-sac
   - argue that the fact that so much hype is being given to these language models is concerning because they are clearly extremely limited
   - argue that a major promise of computational creativity is the augmentation of human creativity, but relying on tools such as ChatGPT will likely have the opposite effect
   - discuss that a well-known science fiction publisher recently closed submissions because of so many low-quality, obviously AI-generated submissions
   - argue that it is likely that the hype has far exceeded reality and we will soon tip into the latest AI winter

[1] The raw output from interacting with ChatGPT can be seen here: `https://tinyurl.com/2p987tuf`

[2] The full paper text was too long to be accepted as a prompt.

[3] I tried to do a minimal amount of this to maintain the "essence" of the writing, though somewhat heavier editing was necessary in spots.

[4] Excluding abstract and title generation.

6. Abstract
   - Write a 100-200 word abstract for a paper that follows the following outline: (followed by all unornamented prompts)
   - Make it much shorter and more high-level (after a very poor first response)
   - More abstract—don't restate lines from the outline (after a second still unsatisfactory response)
7. Title
   - Suggest a title for a paper that follows the following outline: (followed by all unornamented prompts)
   - Suggest something very different, something less literal and more cheeky (after the first several responses were too boring/literal)

Analysis/Discussion

Results The final result is passable as a position paper. It is a correctly, if blandly written, explicitly unsophisticated argument for its stated position; its sophistication, if it has any, is implicit in that fact (and what it implies). While the argumentation is not sophisticated, it is fairly coherent, with both judgements due in large part to the prompts and some post-generation word-smithing/editing (though ChatGPT does get credit for correct spelling, grammar and much of the local cohesion).

Quantitative observations For the five Introduction prompts, ChatGPT generated 1274 words;[5] for the six Understanding prompts, 1243 words; for the five Skill prompts, 1446 words; for the five Dialogic prompts, 1286 words; and for the four Cul-de-sac prompts, 998 words. The resulting initial draft of the full paper text was therefore 6247 words. After editing/proof-reading, I compressed this to 4356 words.[6] For comparison, all 25 prompts combined are 507 words, and this number is even smaller, at only 432 when the content-agnostic ornamentation is removed.

For the abstract prompt, ChatGPT originally generated 306 words (even though the prompt specified a 100-200 word length). This original attempt at an abstract was divided into five paragraphs that was just a clunky summary of the five sections of the paper. As a result, I tried a second follow-on prompt (see prompt outline) asking for more abstraction and shorter length, and repeated this prompt in a slightly modified form as a third prompt before the response was useful enough to work with.[7] This better version was only 150 words long, and I only compressed it to 146 words (though I did additional word-smithing as well).

Process notes Unsurprisingly, initial attempts to have the system write a full paper from a short, high-level prompt[8] were abortive: it seems incapable of counting words, or even accurately estimating how many words it is producing (though perhaps I just didn't use the right prompt to elicit this behavior); it seems incapable of producing responses longer than a few hundred words (possibly less than 700?); and, somewhat surprisingly, it even sometimes refused to argue the case.[9] As a result, I settled on the strategy of "leading" it to make the argument one prompt at a time.

In writing a paper with a new student, for whom it is their first time producing such an artifact, it is often the case that the advisor comes up with the main ideas, sends the student off to expand on them, and then (iteratively) proof-reads and edits the result. The general process here was in some ways quite similar. By contrast, it bore no resemblance to the process of co-authorship with an experienced student, let alone another colleague—there was no exchange of ideas, no change of plan, no shared vision, no excitement, no argument, no mutual understanding and no learning.

In general, the writing produced by ChatGPT was technically correct,[10] but the writing was choppy, repetitive, full of filler and immature, with simple sentence structure, weak, repetitive transitions transitions and a formulaic structure.[11]

While all generated responses were generally somewhat repetitive/redundant, those generated for the introduction were particularly guilty in this regard. ChatGPT seems to see all the prompts as sort of asking the same thing, though they are not. Still, the result was usable as an introduction; it just required significant editing. Of the 25 content prompts used, only three were modified to elicit a better response (based on experience using the system), and only three times was it deemed necessary to ask the system to regenerate a response to any of the 25 prompts. Notably, none of the prompt changes or regenerations were made in a way that changed the narrative (as one would hope might happen in a co-creative environment); they were only changed due to unacceptable output.

Whenever a response seemed to require documentation for a claim, I prompted the system to "provide references to support the preceding response." In many cases, I used the references it provided as close to where they were provided as possible, but I didn't always use all of them, and I occasionally used one somewhere else that I thought was a better fit. I did not augment the references with others, even when it seemed like I should. Notably, the majority of the references supplied by the system were fictitious (and these are noted in the bibliography)—a mashup of complete fabrications, real author names, real venue names, plausible sounding titles and metadata, many broken URLs (and some that point to completely different, unrelated papers).

[5]This does not count the conclusion paragraphs with which it ended most responses, all of which I discarded immediately.

[6]Some of which were mine, so this is a conservative estimate as a compression ratio.

[7]Even with three progressive prompts, I still had to regenerate multiple times.

[8]E.g., write an 8000 word research paper arguing that language models cannot be co-creative partners.

[9]E.g., the prompt "argue that statistical language modeling is a dead-end approach to language *understanding*" produced a response that explicitly refused to do so and instead argued the opposing view; its argument for the counter-position was poor, and, indeed, sort of made the case for lack of understanding I originally requested, though unintentionally.

[10]In the sense of grammar, spelling, language usage, etc.

[11]To be fair, some of this may be due to the ornamentation requiring "formal language".

Of those that were not fictitious, some were actually apropos and useful, while others were somewhat tangential.

Positives The experience was not without positives:

- It was interesting and a bit satisfying to see the paper "grow" before my eyes, like watching a crystal garden as a child.
- Title generation turned out to be the star of the process. For me, coming up with a good title is always difficult, so this was a nice surprise, and something I may actually use in the future. With the first prompt, the system generated boring, literal titles (though perfectly useful, as well, it you like that sort of title); the additional prompt (see prompt outline above), however, was very effective at soliciting many interesting results, and enough genuinely good options to make me suffer over which to choose (a new kind of suffering over the title!) Here are three examples I had a hard time not using:
  - *The Overhyped Co-Author That Can't Even Draw a Stick Figure*
  - *How ChatGPT and its Friends Became the Kardashians of AI*
  - *How AI Language Models are the Bad Dates of Co-Creative Partners*

Negatives The negatives likely far outweigh the positives:

- The general writing quality is poor, as discussed above
- No potential for co-creativity
- There is no way to guarantee that generated "facts" are, in fact, facts, and, therefore, they are very likely not. E.g., the claim in the introduction that ChatGPT currently tops the SuperGLUE benchmark leaderboard with a score of 89.9 appears to be untrue (and, further, appears likely to have never been true).
- Many generated references are partially or completely fictitious, as discussed above
- The process of producing, verifying and including references[12] was painful, but not completely useless. At this point, both generated references as well as more general types of fact cannot be trusted; however, as bad as the situation is at the moment, it is a problem that will likely be remedied by incorporating some types of symbolic knowledge into these models (an ongoing topic of research), resulting (eventually) in a better calculator/theorem prover. For the moment, though, a couple of particularly egregious examples (of many):
  - The Harvard Business Review "reference" (Khosla 2020) on the future of creative work (generated verbatim for two different queries) is especially insidious because there is a real reference, in the Harvard Business Review in fact, that is almost identical.[13]
  - The "reference" about a science fiction publication combines a truth (they really did stop accepting submissions because of an influx of AI-generated garbage[14]) with a bogus reference (Heath 2022) to support it.
- Finally, as an illustration of how these models are superficially sophisticated but fundamentally superficial, an unedited conversation I had with ChatGPT:

  Me: break a grammatical rule to make a point

  ChatGPT: Me not knowing proper grammar don't mean I can't communicate effectively.

  Me: what grammatical rule did you just break?

  ChatGPT: I used "don't" which is a contraction of "do not" with "me" as the subject, but the correct subject pronoun for the third-person singular subject should be "I". So, the correct form would be "I don't know proper grammar", but I used "Me not knowing proper grammar" instead.

  Me: You actually broke two rules and you didn't identify either of them: "Me" should have been "My" and "don't" should have been "doesn't"

  ChatGPT: You are correct. I apologize for my mistake. The correct sentence should be: "My not knowing proper grammar doesn't mean I can't communicate effectively." I broke two grammatical rules: I used "Me" instead of "My" to indicate possession. I used "don't" instead of "doesn't" to agree with the third-person singular subject "My not knowing proper grammar".

Conclusions This experience makes it clear (at least to me) that (statistical) large language models can be a useful tool but are not now nor likely ever to be a candidate for co-creative partner. They are, by their nature, limited in the types of writing/language/artifact they can produce, and that nature precludes both understanding and creativity.

An unintended consequence of this experiment is the implication that at the level of the main paper, the contribution is really just the prompts, with everything else essentially fancy filler. This leads immediately to the possibly uncomfortable question of how much similar (human-generated) filler traditionally produced research papers contain.

Finally, it seems important to reiterate here the warning about the current hype around these large language models (as well as that around other recent AI advances in vision, text-to-image, etc.) potentially leading to the next AI winter. On a brighter note, this sobering possibility presents an enticing opportunity for computational creativity, if the field can avoid being overwhelmed by the current hysteria.

[12] This was the longest, most tedious part of the entire process and would have been even worse if I'd done it entirely properly.

[13] see https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work

[14] The publication was Clarkesworld, and a real reference is here: https://www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence. The story is only four days old at the time of this writing.

Beyond Prompts:
Exploring the Design Space of Mixed-Initiative Co-Creativity Systems

Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, Mark Riedl
School of Interactive Computing, College of Computing
Georgia Institute of Technology
Atlanta, Georgia, USA
{zhiyulin, ehsanu, rohanagarwal, sdani30, vvashishth3}@gatech.edu, riedl@cc.gatech.edu

Abstract

Generative Artificial Intelligence systems have been developed for image, code, story, and game generation with the goal of facilitating human creativity. Recent work on neural generative systems has emphasized one particular means of interacting with AI systems: the user provides a specification, usually in the form of prompts, and the AI system generates the content. However, there are other configurations of human and AI coordination, such as co-creativity (CC) in which both human and AI systems can contribute to content creation, and mixed-initiative (MI) in which both human and AI systems can initiate content changes. In this paper, we define a hypothetical human-AI configuration *design space* consisting of different means for humans and AI systems to communicate creative intent to each other. We conduct a human participant study with 185 participants to understand how users want to interact with differently configured MI-CC systems. We find out that MI-CC systems with more extensive coverage of the design space are rated higher or on par on a variety of creative and goal-completion metrics, demonstrating that wider coverage of the design space can improve user experience and achievement when using the system; Preference varies greatly between expertise groups, suggesting the development of adaptive, personalized MI-CC systems; Participants identified new design space dimensions including scrutability—the ability to poke and prod at models—and explainability.

Introduction

The wider availability of generative AI systems in domains ranging from text (Brown et al. 2020), image (Ramesh et al. 2022), program code (Chen et al. 2021), to game stages and concepts (Khalifa et al. 2020), is making the development of creative content more accessible to people with more diverse backgrounds and skills. Recent work on neural generative systems has emphasized one particular means of interacting with AI systems: the user provides a specification (e.g., prompt, previous text context, structured data, or one work of art to be restylized into another), and the AI system generates the content. However, the means of initialization, along with the majority of interactions between user and system, are *not* human-centered. In particular, they impose a specific paradigm of input on the human designer that best suits the underlying algorithms and models instead of the needs of the human designer.

Other configurations of human designer and AI creative system are possible that promise to reduce cognitive load, frustration, system abandonment (Sweller 2011), and make these systems more casual and enjoyable (Compton and Mateas 2015). *Mixed initiative* (MI) systems are those in which both human and AI systems can initiate content changes.

Co-Creative (CC) systems are those in which both human and AI systems can contribute to content creation Mixed initiative co-creative (MI-CC) systems have been demonstrated in game design (Liapis et al. 2016), coding (*Github CoPilot*), drawing (Davis et al. 2015) and storytelling (Alvarez, Font, and Togelius 2022) Although the building of systems that make use of MI-CC traits may help us better understand how users think and collaborate with creative AI systems, our understanding of the human factors that underly successful MI-CC systems remain relatively under-studied compared to the development of new MI-CC systems.

In this paper, we build on the dimensions of MI-CC systems identified by Lin et al. (2022): Human vs. Agent-initiated, Elaboration vs. Reflection, Global vs. Local. This framework defines a hypothetical design space for MI-CC systems where each value of each dimension can be instantiated as a specific way for a user to communicate creative intent with an AI system (and vice versa). In the domain of story generation, we conduct an exploratory human-participant study with 7 unique MI-CC systems as probes, each representing a plausible subset of the design space. [1] We measure the perceived support the tool affords via the Creative Support Index (CSI) (Cherry and Latulipe 2014). Our study indicates that more extensive coverage of the design space can improve user experience and perceived creative achievement. We also observe that preference for different types of communication with (and from) the system varies with expertise, suggesting the potential for adaptive, personalized MI-CC systems. Finally, our human-participant study uncovers a 4th dimension: explanation.

[1]Code for the systems used in the study is available at `https://github.com/eilab-gt/beyond-prompts-experiment`

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

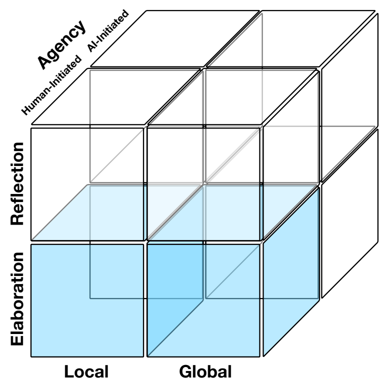


Figure 1: The design space of user-AI communications in mixed-initiative co-creative systems considered in this work, consisting of three dimensions. The blue cubes have been explored in prior studies (Lin, Agarwal, and Riedl 2022).

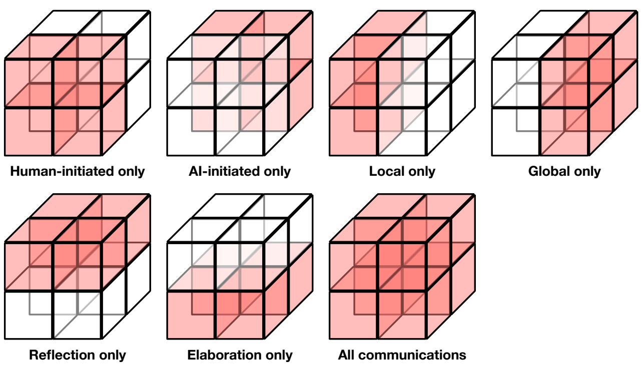


Figure 2: The full system and six ablations—each removing one dimension from the design space—used as conditions in this work.

Background and Related Work

A *mixed-initiative system* is one where "a human initiative and a computational initiative" cooperate towards a shared goal (Novick and Sutton 1997). In this work, we focus on storytelling in a game design setting. However, as Lai, Leymarie, and Latham (2022) point out, it can be easily applied to other creative domains. Like that work we exclude so-called "fire-and-forget" systems (Ramamurthy et al. 2022) from consideration and focus on systems that allow iterative improvement on the creative artifact.

A *co-creative* agent (Rezwana and Maher 2021; 2022; Guzdial and Riedl 2019; Grabe 2022; Kreminski and Mateas 2021) is one that AI "possesses the ability to alter the creative work equal to a human counterpart". Note that when it comes to capability, "ability" does not imply *human-parity*. Furthermore, as the human user and the AI system may tackle different parts of the creative work, it also does not entail equal *responsibility*; Note that a mixed-initiative system does not necessarily need to be co-creative as the final product of the process does not have to be a creative artifact (Sekulić, Aliannejadi, and Crestani 2022).

To study the information that flows between the user and the MI-CC agent, we depart from the definition of Communications from (Lin, Agarwal, and Riedl 2022), which is itself based on works in categorizing or differentiating between different types of interactions between parties in an MI-CC system. The framework by Rezwana and Maher (2021; 2022) that models interactions in co-creative systems that "focus on flow of information between collaborators", inspiring the definition of Communications. Guzdial and Riedl (2019) point out that human designers and AI "can initiate the same action sets to modify the creative work, albeit with different executions". Kreminski (2021) presents a survey on communications "where the agent thinks about what happened in the process and takes actions based on it".

Recent works on co-creative systems with a large array of interactive options include *CoAuthor* (Lee, Liang, and Yang 2022), wherein a text continuation setting fine-grained keyboard-based actions are recorded as a dataset, and *CoPoet* (Chakrabarty, Padmakumar, and He 2022), a poetry system wherein options are implemented as prompts constructed to represent requests with different configurations and ranges of application. These works have shown the potential of co-creativity systems with a wide range of capabilities, further encouraging us on a comparative study over the design space of these capabilities, which is the focus of this work.

Design Space for MI-CC Systems

MI-CC systems can come in a variety of sizes and shapes. Lin et al. (2022) presented a framework to help categorize them in terms of how the user and system communicate with each other and how information flows chronologically from one to the other. The framework is domain agnostic, though demonstrated through a text generation example. The framework contains three continuous, non-exhaustive dimensions to classify communications to and from user and system:

- **Human-initiated vs. Agent-initiated**, which considers which of the two parties is initiating communication.
- **Elaboration vs. Reflection**, which deals with whether the communication relates to previously generated contents (reflection) or newly planned actions (elaboration).
- **Global vs. Local**, which is based on the scope of the creative work of the communication.

We use the framework to guide the construction of variations of a MI-CC storytelling system. By tying an axis A to each dimension, the Cartesian product of these axes form a *design space*. In the design space there are $2^3 = 8$ different ways that user and AI system can communicate. For example, one means of communication might be human-initiated, involve elaboration of the creative artifact, and focus on local information.

The framework—and our work accordingly—considers *social* communication out of scope. Social communication is that between two creators that does not involve the sharing of information about the creative artifact or the alteration of it. Social communication, however, is not entirely unproductive as it may improve the relationship between human

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

creator and AI agent, allow the user to better understand the AI system, and improve trust and rapport (Margarido et al. 2022). The framework is also agnostic about the modality (text, visual, audio, etc.) of the communication, which Margarido et al. (2022) also argues may be considered an additional dimension.

The framework and design space does not tell one *how* user-AI communication should occur beyond the broad properties provided such as who initiates, the extent of the information, and whether the information is about newly planned actions or previously made actions. To that end, each type of communication can be implemented in many ways within a given MI-CC system. In the next section, we provide our specific instantiations of each of the eight types of communication.

Although conditional generation systems—GPT (Brown et al. 2020) for text, but also existing for other modalities (cf., (Ramesh et al. 2022; Khalifa et al. 2020))—condition their generation on input such as a prompt from the user, the choice of information exchanged among candidates is under-explored beyond so-called "fire-and-forget" systems which are trained on text corpus and/or task-based human feedback (Ramamurthy et al. 2022). Fire-and-forget systems, which can also be thought of as *assistants*, are a special case of the design space: human-initiated, elaborative, and local. Studying MI-CC systems with Communications from different points in the design space will help researchers to better understand how an AI system will collaborate with a human designer and facilitate generation systems to better align with tasks unique to MI-CC systems.

Experimental System Overview

To conduct an exploratory study of how the availability of different means of communication affect the actual and perceived creative experience, we used the *Creative Wand* framework (Lin, Agarwal, and Riedl 2022), which is designed to facilitate experimentation with MI-CC systems. The Creative Wand framework is a highly configurable MI-CC system made up of four abstracted components:

- *Creative Context*: an interface between generative algorithms related to the specific domain.
- *Experience Manager*: responsible for maintaining the state of the system.
- *Communications*: a set of modules that instantiate different means of communicating creative intent (see Section Communications.
- *Frontend*: defines how information is presented to, and received from, the user.

Creative Context: Storytelling Domain and AI Algorithms

Similar to Lin et al. (2022), we also consider textual story creation as the domain, situated as a key task in game development. In the story creation domain, the user attempts to create a plot with 10 lines. Since it is a plot, the lines express the general activities of characters. See Figure 3 for an example. Since story creation can be open-ended, we needed a way to constrain the activity in order to assess user performance. To that end, we artificially provide the user with a goal. We gave the same goal that (Lin, Agarwal, and Riedl 2022) used, which is to create a story that starts with "Business" and ends with "Sports" while mentioning "soccer".

In a MI-CC story creation task, the system must be capable of receiving communication from the human designer about creative intent at various levels (global, local), but also providing critical reflection on the story content. As there is no one AI story generation system capable of doing everything we need for all aspects of the design space, We deployed two existing AI systems: *Plug and Blend* (Lin and Riedl 2021) updated to use the larger GPT-J (Wang and Komatsuzaki 2021) language model instead of GPT-2 (Radford et al. 2019); and *CARP* (Matiana et al. 2021).

Plug and Blend This system uses two models to generate text that adheres to topic controls. The first model is a standard, unaltered large language model. In the case of this paper, we use the GPT-J pre-trained large language model, which accepts a prompt or context text that begins the story. The second model learns a set of weights that can be applied to the output logits of the language model output in order to bias the generation toward a particular topic. A set of topics and the sentence spans they should be applied to is provided as a second type of input to Plug and Blend, referred to as a *sketch*; They are translated to individual control strength that amplifies the weight applied from the second model, and further guides the generation of paragraphs. We modified the pipeline so that "Story for kids: Once upon a time,", concatenated with at most two previous lines, as the prompt, along with the topic control, is used to generate each line of the story in our system.

CARP We used the CARP model (Matiana et al. 2021), a language model that is trained on contrastive objectives to learn a cosine similarity score between a sentence and a short critique, such as "This character is confusing". CARP cannot generate narratives, but can score individual lines in a narrative according to a given criterion. It is the basis for our communication modules involving reflection. CARP produces values between $[0.15, 0.4]$, which we rescale to $[0, 1]$.

Communications

We designed 11 modules for communication to cover the entirety of the design space, as well as some additional communication modules that emulate basic functionality that many expect in creation tools. We give the 11 communication modules below, indicating where on each of the three axes it falls. We organize the list around the axis of elaboration vs reflection because elaboration and reflection are tied to the two AI algorithms.

Elaboration Communication Modules Elaboration communications are related to generation of new contents, and use the Plug and Blend AI algorithm. To batch user input for a better user experience, we do not immediately start the regeneration process until the user requests the story to be rewritten, which is one of the three miscellaneous communications.

- **Write a sentence** $\left[\begin{smallmatrix}Elab.\\ Human\\ Local\end{smallmatrix}\right]$ The user provides a specific sentence to be inserted at a particular line index. If there is already text in that line, it gets replaced.

  Example: The user replaces the first line with "Hello!".

- **Apply a topic** $\left[\begin{smallmatrix}Elab.\\ Human\\ Global\end{smallmatrix}\right]$ The user provides a topic code along with a starting line index and an ending index. We provide four pre-defined options—"business", "sports", "science", "world". The user can type in their own free text code as well. The Plug and Blend sketch data structure is updated though generation does not happen until the user requests re-generation.

  Example: The user applies "Business" to the first five lines of the story.

- **Get a sentence suggestion** $\left[\begin{smallmatrix}Elab.\\ Agent\\ Local\end{smallmatrix}\right]$ The Plug and Blend AI algorithm generates a new candidate sentence based on a random existing line and a topic chosen between "Business" and "Sports", focusing in on the two goals participants are asked to meet. The user can then choose to accept or reject the suggestion.

  Example: The agent provided "Football is interesting..." as a suggestion to line 3 of the story.

- **Get a topic suggestion** $\left[\begin{smallmatrix}Elab.\\ Agent\\ Global\end{smallmatrix}\right]$ The user is provided a topic suggestion between "Business" and "Sports" chosen randomly. If the user decides to accept the suggestion, they continue the process as in applying a topic control.

  Example: The agent provides "Sports" as the suggested topic. The user accepted the topic and decided to apply it to the last 5 lines of the story.

Reflection Communcation Modules This family of communications uses the CARP model, which pairs a critique and each sentence with a score signifying how related they are. For each communication below, this information is used differently, but all towards providing the user with insights or opportunities to think about whether the story so far needs further modification. Being provided with a critique, the system highlights lines of the story with increasingly brighter color relative to the rescaled score.

- **Off-topic checker** $\left[\begin{smallmatrix}Reflect\\ Human\\ Local\end{smallmatrix}\right]$ The user picks a sentence in the story and gives a topic, and the agent tells the user whether it's related to that topic. We used "This part of the story should be related to $\langle input \rangle$" as the critique for the CARP model.

  Example: The user selects line 3 which says "Football is interesting..." and asked whether it is related to "Science".

- **Reflect together** $\left[\begin{smallmatrix}Reflect\\ Human\\ Global\end{smallmatrix}\right]$ The user gives a critique, and the agent highlights sentences based on the score given from the CARP model.

  Example: The user provides "It should be raining", and the agent highlights a line in the story that says "It's a sunny day" (as well as any other lines based on how much they fail to match the critique).

- **Get a local story quality tip** $\left[\begin{smallmatrix}Reflect\\ Agent\\ Local\end{smallmatrix}\right]$ The agent picks a tip from a list of pre-determined critique prompts and highlights sentences based on the scores based on how much the match the critique according to CARP.

  Example: The agent picks the "The story should be fun" pre-defined critique and highlights line 7 of the story, showing that the agent thinks that line is fun.

- **Get a high-level story quality tip** $\left[\begin{smallmatrix}Reflect\\ Agent\\ Global\end{smallmatrix}\right]$ The agent picks a tip from the same pre-defined set of critique prompts and tells the user whether the story as a whole is related to it.

  Example: The agent picks "Whether the story is about Sports" from its set of critiques and tells the user "yes", confirming the user that the story is about "sports".

Miscellaneous Communication Modules We also included three additional communication modules to fill in functionality of the system.

- **Write the whole story** When selected the system will regenerate all lines in the story, given a context prompt and the Plug and Blend sketch. This communication can be used multiple times in succession for alternatives.
- **Rewrite from a specific line** Instead of starting fresh, the system only generates lines for the story after a specified point, leaving previous lines intact. This can only be used if a story already exists.
- **Undo** The system reverts the last operation.

Experience Manager and Frontend

We implement a turn-based experience manager based on the sample Creative Wand implementation in Lin et al. (2022). For each turn, the manager provides available options for the user in a chatbot-like dialogue box, each of which maps to a communication module (for agent-initiated communications, entry point to allow the agent to take the initiative.) See Figure 3, which shows a portion of the user interface. We extend the Creative Wand framework to provide user experience by enhancements such as reverting back to previous states (i.e., undo).

Study Methodology

To study the design space of communications in MI-CC systems, we developed seven versions of the story creation system described in the previous section. One version had communication modules from every part of the design space. The other six versions removed communications along a single dimension. See Figure 2.

In this exploratory study we seek to determine how the presence or absence of different modes of human-AI communication affect perceptions of creative support. We also seek to determine if individual variables such as creative background and familiarity with AI affects the above.

We recruited 185 participants on Prolific[2] who were automatically screened by the platform for adequate English proficiency. Each experiment session lasts for approximately 30 minutes, and we paid the participants $15 per hour.

Participants are first asked to complete a questionnaire about their creative background and familiarity with AI. There were three multiple choice questions, as given below with their possible choices:

[2] `prolific.co`

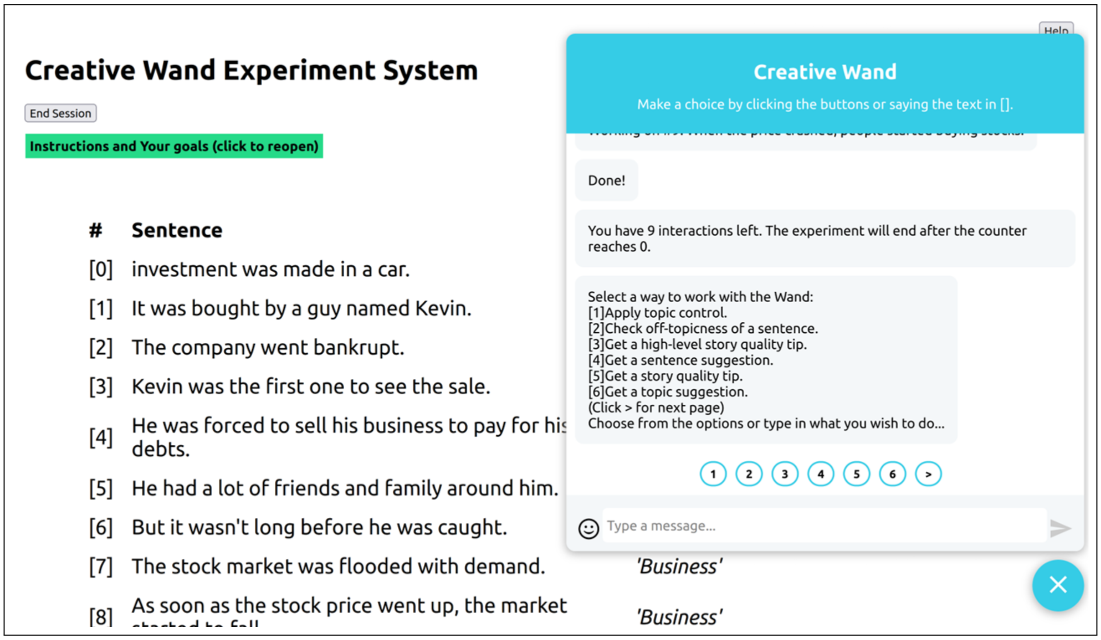


Figure 3: Screenshot of our experiment system, along with instructions.

A. Level of confidence with using a computer to author contents:
   A1. I do not use computers to create things.
   A2. I have used computers to create things, but for the past year, I have not done it once a week.
   A3. I use computers to create things more than once a week, but I'm doing it not for the job (for example, for interest).
   A4. I use computers to create things for my job.
B. Level of confidence with using a computer to create games:
   B1. I never used a computer to create anything related to games.
   B2. I've done some work in the realm of games, but for the past year, I have not done it once a week.
   B3. I create content for games out of interest, for more than once a week.
   B4. I create content for games for my job.
C. Familiarity with AI:
   C1. All I know is no more than it being a buzz word.
   C2. I have experience using something with "AI technologies" with it.
   C3. I understand how recent AI technologies work.

We then show instructions to familiarize them with how to use the experiment system. This consists of annotated screenshots of the interface during different stages of the study, and a brief introduction to the workflow of co-creating a story.[3]

[3]Individual communications are not included in the tutorial, but further instructions are triggered the first time each communication is activated.

Participants are then assigned to a random condition in which they will interact with two versions of the system: the "full" system (all communication modules), and one of the six ablations (communication modules removed in one dimension). We counter-balanced the order we present the systems so that participants randomly start using either "full" or the ablation we assigned.

For both systems, participants were asked to create a story that was 10 lines long, starts with the topic of "business", ends with the topic of "sports", and mentions "soccer" at least once. This is the same task and goal criteria as used by Lin et al. (2022). Participants are given 12 interactions with each system. An interaction is only complete once the participant provides all information for the system to execute the option and doesn't change or cancel the communication.

Once they finished using both systems, as the exit survey, participants were asked to complete another questionnaire with seven questions about their satisfaction with the process and the generated story. These questions are presented in random order for each participant. The first six questions are adapted from the Creative Support Index (CSI) (Cherry and Latulipe 2014), which is a validated measure of how well a tool supports human creativity. We ask: between the "full" system and the ablation, "Which system is more associated with each of the statements?":

Q1. **(Expressiveness)** This system made it easiest for me to express and share my goals, given to me in instructions, with the AI system.

Q2. **(Enjoyment)** I enjoyed interacting with this system most.

Q3. **(Exploration)** This system was most helpful for exploring different ideas and possibilities.

	Overall	Agent-Init. Only	Human-Init. Only	Elaboration Only	Reflection Only	Global Only	Local Only
Num. valid responses	185	31	32	30	32	27	33
Q1: Expressiveness	62.2%*	74.2%*	46.9%	56.7%	78.1%*	63.0%	54.5%
Q2: Enjoyment	60.5%*	74.2%*	43.8%	50.0%	81.2%*	59.3%	54.5%
Q3: Exploration	62.7%*	71.0%*	46.9%	56.7%	71.9%*	70.4%*	60.6%
Q4: Immersion	62.2%*	71.0%*	50.0%	60.0%	75.0%*	59.3%	57.6%
Q5: Collaboration	59.5%*	71.0%*	40.6%	56.7%	81.2%*	59.3%	48.5%
Q6: Result worth effort	60.5%*	64.5%+	53.1%	60.0%	71.9%*	66.7%+	48.5%
Q7: Better responses	61.6%*	67.7%*	56.2%	63.3%	78.1%*	59.3%	45.5%

Table 1: Rate of participants that preferred the Full System over the ablations. * represents a significance level of $p < 0.05$ on Full system preferred over the ablation; + for $p < 0.1$. No ablation was preferred statistically significantly.

Q4. **(Immersion)** This system made me feel the most absorbed in the task to the point that I forgot I was working with the system.

Q5. **(Collaboration)** This system best allowed me to achieve the goal assigned to me.

Q6. **(Results worth effort)** This system provides the overall best quality stories by the time I was done.

Additionally, we also asked:

Q7. Which system tends to get the best response for the same type of requests?

We anticipate no preference between both systems on Q7, as the implementations of how the systems handle these requests (provided that an ablation system has that capacity) is unchanged. As an attention mechanism, we also asked what the perceived similarity and differences between these two systems are before the participant finishes the study.

Results and Discussions

Participant Creative Background 98% of the participants reported that they at least used computers to create things (A2-A4), and 41% say they do it as their job (A4). Although we did not specifically recruit people with experience in designing game contents, 49% of the participants identify them as at least carrying out some work in the realm of games (B2-B4). 84% reported that they have used something with AI technologies (C2-C3), and 26% say they know how recent AI works (C3). Table 2 (column 1-2) shows how many participants responded yes to each question.

Perceptions of Creativity Support Table 1 shows the preference of users between the "full" system and the ablations on the seven questions Q1–Q7. Participants prefer the full system overall. When considering only the Agent-Initiated and Reflection ablations, the preference for the full version is also statistically significantly preferred on all questions. That is: removing human-initiated communication or elaboration communication significantly degraded the creative experience in every measurement. The ability to fully or partially generate the story was always an option.

The Global-only ablation, which removed communications involving local changes, was significantly less preferred than the full version only when considering the questions on exploration, and "results worth it". This suggests that global communications were not sufficient alone for exploring different ideas and participants felt less overall satisfaction with their story results when unable to make localized changes. Even though in many cases the full version was preferred over other ablations more than 60% of the time, when participants are spread across conditions, there is a higher bar for statistical significance.

Participants were the most indifferent when comparing the full system to Local and Human-Initiated ablations, removing Global and Agent-initiated communications, respectively. That is, removing these resulted in less reported degradation of the creative experience. In the Human-Initiated ablation, the AI is the most passive and never does anything until users provide enough information for them. Most non-MI-CC systems operate this way and may be used at least weekly by 75% of the participants (A3-A4). Global communications are likely harder to use than local communications. Participants were asked to learn a new creative support tool in a less-than-30-minutes experiment, with a sharp learning curve toward mechanisms that the participants are not familiar with in the first place. These might have played a role. Although a follow-up longitudinal study may help investigate this effect and provide a more accurate picture to study *specific parts of the design space*, in our opinion, this also hints that not all dimensions in the design space of communications are of equal value to users.

The full system was not significantly preferred over the Elaboration-only ablation, even though the full version was preferred 56%–63% of the time. This suggests that participants were more sensitive to the loss of reflective communications than the loss of agent-initiated or global communications. The role of reflective communication deserves further study; this study cannot determine the extent to which the specific use of CARP as the model for processing story critiques play in participant perception of reflective communications.

Surprisingly, participants prefer the "full" system on Q7—the system provides better responses—despite the fact that the AI systems were the same across all systems (when not removed due to the elaboration-only or reflection-only ablations). We hypothesize that because the "full" system can help the users achieve the goal better (61.6% with statistical significance), it is likely that intermediate stories are also easier to work with; the communications "give the better response" because they have a better story to work on.

The overall trend is that **a wider coverage of the de-**

	n	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Welch's t-test		
Level of confidence with using a computer to author contents (A1-A4)									**A2**	**A3**	**A4**
A1	3	*Too few participants*									
A2	44	61.4%	56.8%	54.5%	54.5%	61.4%	56.8%	56.8%	N/A	*	
A3	62	66.1%	67.7%	72.6%	67.7%	66.1%	69.4%	71.0%	*	N/A	*
A4	76	60.5%	57.9%	60.5%	63.2%	53.9%	56.6%	57.9%		*	N/A
Level of confidence with using a computer to create games (B1-B4)									**B1**	**B2**	**B3**
B1	94	63.8%	63.8%	60.6%	66.0%	64.9%	67.0%	66.0%	N/A	*	
B2	65	52.3%	50.8%	66.2%	52.3%	46.2%	49.2%	53.8%	*	N/A	*
B3	21	81.0%	76.2%	61.9%	76.2%	71.4%	61.9%	61.9%		*	N/A
B4	5	*Too few participants*									
Familiarity with AI (C1-C3)									**C1**	**C2**	**C3**
C1	29	48.3%	55.2%	48.3%	51.7%	48.3%	58.6%	58.6%	N/A	*	+
C2	107	66.4%	64.5%	67.3%	67.3%	61.7%	58.9%	63.6%	*	N/A	
C3	49	61.2%	55.1%	61.2%	57.1%	61.2%	65.3%	59.2%	+		N/A

Table 2: Rate of preference on Full system, grouped by answers to demographics questions. Only data for groups with more than 20 participants are shown. * means different distribution with $p < 0.01$, + for $p < 0.1$.

sign space of user-AI communication types is appreciated. The study provides in-depth understanding of the relative significance of different types of user-AI communication. The fact that our study cannot distinguish a preference for Local and Global communications suggests that they may both be important (except where noted otherwise above). This is notable due to the absence of global communication modes in most "fire and forget" systems.

Individual Differences For each background experience question (A, B, C), we place participants in separate groups based on which multiple-choice option they selected. This creates groups A1 through A4, B1 through B4, and C1 through C3. Table 2 shows the number of participants sorted into each group and how they responded to the creativity support questions. We excluded groups with less than 20 participants from the analysis.

We conduct a Welch's t-tests between all groups and observe strong significant differences ($p < 0.01$) between Groups A2 and A3, A3 and A4, B1 and B2, B2 and B3, and C1 and C2 with regard to how each group responded to the creativity support questions. Except for C3 in which we observe a weaker difference ($p < 0.1$) between C1. Participants from each expertise group differ in their preference for the full system from at least one other group, suggesting that **MI-CC tools should be customized to different types of users with different levels of creative expertise and AI familiarity**.

Creativity Support Index Questions are Correlated We conduct a correlation analysis on the questions asked in the survey (Figure 4). We observe a medium (0.43) to strong (0.79) correlation between the six questions adapted from the Creativity Support Index (Cherry and Latulipe 2014). While we would expect all questions to correlate with each other because they are all, at some level, measuring different dimensions of creativity support, the data reflects participant response to the presence or absence of communication types. We see Q1 (expressiveness) and Q2 (enjoyment) and

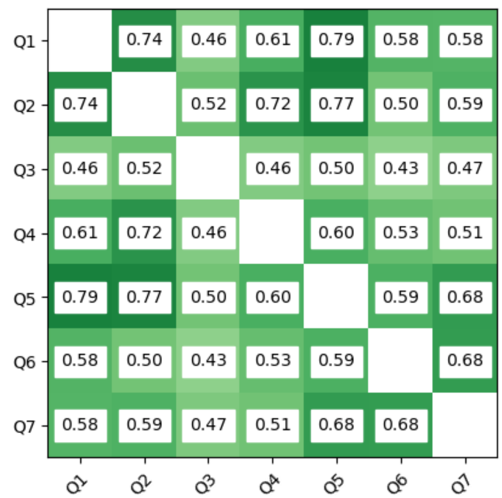


Figure 4: Correlation between questions in the survey.

Q5 (collaboration) correlate at > 0.7, suggesting the factors impact user perceptions of expressiveness are the same that impact perceptions of enjoyment. Likewise we see Q2 (enjoyment), Q4 (immersion) correlate at > 0.7, suggesting that the factors that impact perceptions of joy are the same that impact perceptions of immersion. Due to the nature of the task, these factors are the presence or absence of certain communication modes, though we do not have fine-grained detail enough to identify which ones. This further suggests that **the communications that make the creative experience enjoyable are the same as those that make the experience seem immersive, expressive, and collaborative**.

Q7 (better responses) is strongly correlated with Q6 (results worth it). Q7 was not derived from the Creativity Support Index, but this correlation provides further explanation for the observations about Q7 earlier that the perceived quality of AI responses would be correlated with perceptions of satisfaction with the creative outcome.

Qualitative Findings & Discussion

We also analyzed open-ended justifications participants provided for their perceived levels of satisfaction with the systems using approaches inspired by thematic analysis (Aronson 1994). Taking an inductive approach, we started the process with an open-coding scheme and iteratively produced in-vivo codes (generating codes directly from the data). Next, we analyzed the data using axial codes, which involves finding relationships between the open codes and clustering them into different emergent themes. Through an iterative process performed until consensus was reached, we share the most salient themes below.

Participants valued the **ability to exercise control** over the co-creative writing tool. Whichever tool was "easier to control the topic" (P09) was often favored. Customizability was a prized asset– they felt that the customizability allowed "the story to go together" and be more coherent (P98). This notion of controllability was also associated to the tool's ability to "take topics into consideration" (P64), which indicates how participants ascribed comprehension abilities in the tool based on their ability to control it. The sentiment is expressed succinctly by the following participant:

> Fish[4] is superior to Rabbit[5] in that it you can guide and interact with it and *it listens to feedback* and doesn't just write what it wants. Fish allowed you more control in guiding the story on topics before starting so it was more accurate and also more customizable. Rabbit felt more random with less options and control, it started off topic and stayed off topic even when being prompted. Fish overall was a lot better than Rabbit. (P29, emphasis added)

There was a **desire for scrutability**—to poke and prod—to **get a mechanistic or functional understanding** of the tools. This theme follows from the previous theme around the desire to control. Participants exhibited a desire to "understand the mechanism of checking how one sentence is related to a particular topic" (P07). The more control a tool allowed, the higher its perceived scrutability. Participants were trying to achieve a mechanistic understanding (Lombrozo and Wilkenfeld 2019)—how things worked—as well as functional understanding (Lombrozo and Gwynne 2014) of the "why" behind the actions. A major part of this understanding was *reciprocal* and *mutual*; that is, participants felt that they could understand the tool if the tool could understand their instructions or input:

> I had an easier time understanding the Fish system. And it appeared to understand the topics better based on my interaction. (P76)

A core implication of both of these findings around control and scrutability suggest that adding explainability to these systems can enable argumentations, expose creative processes and augment the user's mental model (Llano et al. 2022) and thereby foster better collaboration. Emerging work in Explainable AI (XAI) showcases that user backgrounds matter—that is, who opens the "black-box" matters when it comes to making sense of the AI's output (Ehsan et al. 2021b). This entails that we need to customize the explainability according to the user's background (which can include AI literacy, levels of experience, and familiarity). Moreover, we can also fine-tune explanations that target specific types of understanding such as mechanistic or functional. Each type of understanding is goal dependent; therefore, the explanations also have to be appropriately actionable (Ehsan et al. 2021a).

[4]Codename for Full system

[5]Codename for the Reflection-only ablation.

Limitations

While we aim at studying the design space of communications by picking up ones that best express their neighbourhood, the three dimensions we borrowed from Lin et al. (2022) is not complete; communications can also feature traits from both sides of an axis, such as "adding details to an existing sentence" being both elaborative and reflective. As we focus on **what** information is passed between the user and the AI agent, we controlled all the systems we used in the experiments to use the same User Interface and limited representation of text and highlight colors. We invite colleagues alike to conduct similar experiments on other dimensions of Communications and representations, potentially in other modalities (image, speech, and more).

Arguably, agent-initiated communications still need human users' approval to initiate, as the particular implementation requires all communications to be triggered by the user selecting an option in the menu. We made this decision to unify the representations of the Communications in the study. Although we argue that the capability of the agent selecting which communication to trigger *actively*, which is ultimately a decision-making problem over all communications, a topic that is beyond the scope of this work, Since the goal of our work is to study MI-CC systems, we decided to pick a generative system that strikes a balance between availability and consistency with regard to the MI-CC experience we need for the study.

Discussions

Generative language models (LMs) rapidly advance - While this paper was being reviewed, ChatGPT, GPT-4[6] and a family of large-scale LMs that utilizes RLHF (Ramamurthy et al. 2022) demonstrated to the whole world end-to-end capabilities for collaborative authoring, where a dialogue agent can both generate contents based on an initial prompt, and amend what is just generated with follow-ups, all provided by human users in natural languages. The usage of RLHF, where a reward model is trained to forecast human preference for the dialogues and then used to influence what LMs generate, is a crucial asset of these systems with regard to MI-CC. However, such systems are nominally mixed-initiative, as they by design is a question-answering and continuation assistant, by design only providing *post-hoc* [7] con-

[6]chat.openai.com

[7]*Post-hoc* prompt-based explanations answer "Based on the decision I *already* made, why?". Due to their token-based probabilistic continuation nature they are not designed to give *ad-hoc* explanations of how they made decisions.

tents and explanations when user requests them; They also have spaces for improvements as a co-creator, as RLHF-enhanced LMs rely solely on the context (LM) and "mean" preferences of *a sampling of general public* (RLHF), which is insufficient as we already demonstrated in this work that at least user-specific preferences and their prior experience also plays a role.

We believe, with these LMs and alike showing generation capability *when the prompts are right*, the golden age of studying MI-CC systems has arrived: **Beyond prompts**, MI-CC systems that stands on these new frameworks have the potential to learn how to collaborate with specific users and truly co-create contents without the cognitive load of prompt engineering and procedures alike. "Instead of a model teaching *you* how to work with it, you should teach and collaborate with it." We leave this as future work.

Conclusions

We present a comparative study with 185 participants on MI-CC systems that only differs in their inclusion or exclusion of particular modes of user-AI communication. We find a trend that MI-CC systems with a wider coverage of user-AI communication types is appreciated, and that preference also varies greatly between expertise groups, suggesting for the development of customized MI-CC systems for different types of users. Participants also exhibited a desire for scrutability– to poke and prod–to develop a mechanistic and functional understanding of the system where explanations can be useful.

Based on this evidence, we recommend that designers of MI-CC systems should pay attention to the *design space* of user-AI communications, carefully study their audience, and plan for adaptation of their system towards individual users, when sketching the interaction paradigm. These insights can facilitate further MI-CC research, and, most importantly, encourage tailored collaborative experience for each designer (of diverse experience levels) to achieve their potential during co-creativity as well as the final output of the process.

References

Alvarez, A.; Font, J.; and Togelius, J. 2022. Story Designer: Towards a Mixed-Initiative Tool to Create Narrative Structures. arXiv:2210.09294 [cs].

Aronson, J. 1994. A pragmatic view of thematic analysis: the qualitative report, 2,(1) spring.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; and Askell, A. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33:1877–1901.

Chakrabarty, T.; Padmakumar, V.; and He, H. 2022. Help me write a poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing. arXiv:2210.13669 [cs].

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. d. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder, P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs].

Cherry, E., and Latulipe, C. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21(4):1–25. Publisher: ACM New York, NY, USA.

Compton, K., and Mateas, M. 2015. Casual creators. In *Proceedings of the sixth international conference on computational creativity*, 228.

Davis, N.; Hsiao, C.-P.; Singh, K. Y.; Li, L.; Moningi, S.; and Magerko, B. 2015. Drawing Apprentice: An Enactive Co-Creative Agent for Artistic Collaboration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, C&C '15, 185–186. New York, NY, USA: Association for Computing Machinery.

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021a. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Ehsan, U.; Passi, S.; Liao, Q. V.; Chan, L.; Lee, I.; Muller, M.; Riedl, M. O.; et al. 2021b. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*.

Grabe, I. 2022. Towards a Framework for Human-AI Interaction Patterns in Co-Creative GAN Applications. *HAI-GEN workshop at IUI 2022* 11.

Guzdial, M., and Riedl, M. 2019. An interaction framework for studying co-creative ai. bibtex.eprint: 1903.09709 bibtex.archivePrefix: arXiv.

Khalifa, A.; Bontrager, P.; Earle, S.; and Togelius, J. 2020. PCGRL: Procedural Content Generation via Reinforcement Learning. bibtex.eprint: 2001.09212 bibtex.archivePrefix: arXiv.

Kreminski, M., and Mateas, M. 2021. Reflective Creators. *International Conference on Computational Creativity (ICCC) 2021* 10.

Lai, G.; Leymarie, F. F.; and Latham, W. 2022. On Mixed-Initiative Content Creation for Video Games. *IEEE Transactions on Games* 14(4):543–557. Conference Name: IEEE Transactions on Games.

Lee, M.; Liang, P.; and Yang, Q. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems*, 1–19. New Orleans LA USA: ACM.

Liapis, A.; Yannakakis, G. N.; Alexopoulos, C.; and Lopes, P. 2016. Can computers foster human users' creativity?

Theory and praxis of mixed-initiative co-creativity. *DCE*. Accepted: 2018-04-23T12:31:38Z Publisher: DCE.

Lin, Z.; Agarwal, R.; and Riedl, M. 2022. Creative Wand: A System to Study Effects of Communications in Co-creative Settings. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 18(1):45–52. Number: 1.

Lin, Z., and Riedl, M. O. 2021. Plug-and-Blend: A Framework for Plug-and-Play Controllable Story Generation with Sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 17, 58–65. Issue: 1.

Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2022. Explainable Computational Creativity. arXiv:2205.05682 [cs].

Lombrozo, T., and Gwynne, N. Z. 2014. Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience* 8:700.

Lombrozo, T., and Wilkenfeld, D. 2019. Mechanistic versus functional understanding. *Varieties of understanding: New perspectives from philosophy, psychology, and theology* 209.

Margarido, S.; Machado, P.; Roque, L.; and Martins, P. 2022. Let's Make Games Together: Explainability in Mixed-initiative Co-creative Game Design. In *2022 IEEE Conference on Games (CoG)*, 638–645. ISSN: 2325-4289.

Matiana, S.; Smith, J. R.; Teehan, R.; Castricato, L.; Biderman, S.; Gao, L.; and Frazier, S. 2021. Cut the CARP: Fishing for zero-shot story evaluation. arXiv:2110.03111 [cs].

Novick, D. G., and Sutton, S. 1997. What is mixed-initiative interaction. In *Proceedings of the AAAI spring symposium on computational models for mixed initiative interaction*, volume 2, 12.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1(8):9.

Ramamurthy, R.; Ammanabrolu, P.; Brantley, K.; Hessel, J.; Sifa, R.; Bauckhage, C.; Hajishirzi, H.; and Choi, Y. 2022. Is Reinforcement Learning (Not) for Natural Language Processing?: Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization. arXiv:2210.01241 [cs].

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with CLIP latents. bibtex.eprint: 2204.06125 bibtex.archivePrefix: arXiv.

Rezwana, J., and Maher, M. L. 2021. COFI: A framework for modeling interaction in human-AI co-creative systems. In *ICCC*, 444–448.

Rezwana, J., and Maher, M. L. 2022. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction*. Publisher: ACM New York, NY.

Sekulić, I.; Aliannejadi, M.; and Crestani, F. 2022. Evaluating Mixed-initiative Conversational Search Systems via User Simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, 888–896. New York, NY, USA: Association for Computing Machinery.

Sweller, J. 2011. Cognitive load theory. In *Psychology of learning and motivation*, volume 55. Elsevier. 37–76.

Wang, B., and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. bibtex.note:Accessed: 2022-05-30.

Focusing on artists' needs: Using a cultural probe for artist-centred creative software development

Luís Espírito Santo[1,2], André C. Santos[1], Marcio Lima Inácio[1]

[1]CISUC - University of Coimbra, Polo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal
lesanto, andresantos, mlinacio@dei.uc.pt

[2]AILab - Vrije Universiteit Brussel, 3rd floor, Pleinlaan 9, 1050 Brussels, Belgium
luis.espirito.santo@vub.be

Abstract

One of the ultimate goals of Computational Creativity research is to make novel, better, and useful software that can be used for creative purposes. The new wave of learning-endowed generative systems has highlighted the potential of AI for creative tasks, so demand for creative software development is expected to grow significantly, which in turn entails the need for adapted software engineering techniques. We conducted interviews and used a digital cultural probe that posed as a virtual co-creative companion with unlimited capabilities to collect qualitative data on how creative fellows, from different areas and with no knowledge about generative models, would use an ideal piece of creative software. By following an Inductive Thematic Analysis, we bring forward a set of domain-agnostic patterns of how software can help in creative tasks. These themes - 12 user needs and 8 contexts of use - can be used to organise functional requirements to sustain an improved user-centred development of creative tools, or might even be used as a classification framework for creativity tools and co-creative systems. Finally, we discuss the benefits and limitations of our methodology that can be repurposed for a more suited and artist-centred initial process of functional requirement gathering.

Introduction

Over the course of last year, several learning-endowed generative systems have been put forward such as text2image models (Ramesh et al. 2022), text2video (Singer et al. 2023), text2audio (Agostinelli et al. 2023) and even new conversational agents such as ChatGPT (InstructGPT) (Ouyang et al. 2022) that have been subject of a lot of attention as well as fierce discussions, especially in regard to their use in creative tasks. These discussions usually revolve around topics such as autonomy, originality, authorship, copyright infringement and the potential negative impact these might have on human creators, even leading to artist-led movements against AI-generated products.

Human-centred design methodologies emerged as a way to fight against the development of potentially unethical or detrimental software (Gasson 2003). This kind of methodology places the person at the centre of the design process and has been specifically tailored to make the final product more useful to its users, not only because these processes directly influence the perception of the involved users, but also because they focus on tackling the real problems of a specific class of people. These often imply the early involvement of users during the initial stages of requirement gathering (RG) to study the desired role the software should play.

The aforementioned automatic generation software is not unique in its capacity of helping with creative tasks. Indeed, we hypothesise that this is only suited for a narrow specific purpose in a wide class of problems that creative workers face daily. We define the class of creative software as any digital tool that users find helpful in what they perceive as a creative task. Applying human-centred design methodologies during the development of this kind of software might facilitate its embedding into society in the long-term (Colton et al. 2015). We then honed in on the question: how do creators think software can help in their creative tasks?

Since functional requirements depend on various factors, such as the task, domain, and humans involved, we believe that bypassing a proper RG process goes against the fundamental principle of human-centred design. Therefore, gathering functional needs directly from artists through a bottom-up approach may be a challenging task, as there is limited literature on this approach. Additionally, there is a lack of general guidelines or tools that can support these individual procedures that focus on functional needs, are artist-driven, and are adapted for creative software. For this reason, we chose to follow a more comprehensive bottom-up approach by involving artists to gather domain-agnostic and general needs for creative software and identify common challenges (ex: the issue of confining the creative process). As a result, we formulated a secondary research question: what methodology can help overcome some of the challenges of collecting needs for creative software?

We conducted a qualitative study in which we interviewed people from several creative fields, who made use of a virtual cultural probe to express their needs over the course of a week. Using Inductive Thematic Analysis, we established common domain-agnostic themes of needs and their contexts of use. We propose two use cases for the resulting themes: a grouping framework for functional requirements for creative software; a categorisation of existing creative software, according to their capabilities. We also believe the methodology we followed can also be used to overcome some of the challenges of RG for specific creative software.

The paper is structured as follows: first, we present previous literature on RG, human-centred design, co-creativity, and cultural probes; we continue by detailing our methodology; next, we present the underlying resulting themes and discuss other additional relevant findings while pointing out the limitations of our methodology and how it can be improved and adapted; finally, we summarise our contributions, provide practical use cases for our themes as well as possible directions for future research.

Related Work

The field of Software Engineering focuses on improving software development methodologies. Researchers in this area agree that the development process should start with requirement gathering (RG) (Rodriguez, Wong, and Mauricio 2017), a task studied in the field of Requirement Engineering (RE). Functional requirements are the capabilities desired for a software system to meet the needs of its users and stakeholders. In contrast, non-functional requirements consider other components of interaction such as performance, security and availability. The development of new kinds of software such as distributed and AI systems exhibit particular challenges and benefit from approaches particularly designed for such contexts, such as RE4DIST (Wirtz and Heisel 2019) and RE4AI (Heyn et al. 2021; Ahmad et al. 2021; Pei et al. 2022).

The design of (co-)creative systems is no different, in the sense that it poses its own challenges to the process of RG and elicitation. Not only the new kind of creative systems can make use of different components (distributed data-driven components, real-time or conversational interfaces, or even generative modules) making them very complex systems, but also serendipity is one of the main and most valued characteristics of creative behaviour (André et al. 2009) which becomes one of the biggest challenges when collecting information for designing creativity-support systems since it is hard to confine the creative act to a specific place and time – creativity cannot be scheduled.

Our search revealed that the published literature on software engineering for creative systems is still scarce and that studies focused on RE for (co-)creative systems (RE4CREATIVE) are even rarer; no empirical study on the collection of functional requirements from real potential users was found. Yet, studies show that not only there seems to be "a positive relationship between users' involvement during RE and system success"(Bano and Zowghi 2013) but also that "the most significant user involvement occurs at the beginning of product development"(Kujala 2008).

Meanwhile, Human-Computer Interaction (HCI) has been attracting attention in the area of creative systems, but the methodologies proposed under this branch are usually more appropriate for analysing the interaction between an already deployed model - or at least a prototype - and usually focus more on non-functional requirements and user experience, being of limited use for acquiring functional requirements. This seems to be what Gasson (2003) criticises when advocating for human-centred over user-centred design.

For example, Kantosalo et al. (2020) inspire themselves in the HCI concepts of modalities, styles, and strategies, and adapt them to the context of co-creative systems, in this way developing a framework "to equip co-creativity researchers with a domain-agnostic vocabulary to discuss the capabilities and shortcomings of existing and proposed interfaces for co-creation." (Kantosalo et al. 2020). In another publication, Kantosalo and Jordanous (2021) present several theoretical roles that a computer can play in human-computer creative collaboration. However, while these studies empower researchers with theoretical frameworks for comparing and designing different approaches, they do not directly involve users nor address how to openly collect their needs.

More recently, another similar framework, Co-Creative Framework for Interaction design (COFI), was proposed by Rezwana and Maher (2022a) where the authors take a similar approach, this time by performing a literature review and using concepts from several fields, including Computational Creativity (CC) and Computer-supported cooperative work (CSCW), to inform their thought process and establish interaction models for co-creative systems. They also analysed a total of 92 co-creative systems using COFI, detecting an underutilised space of possibilities in terms of interaction design. But again, while this framework is useful to track and explore the space of interaction design of co-creative systems, the study does not involve users and does not directly focus on their needs nor functional requirements.

Qualitative fieldwork through qualitative research is essential to obtain a deep understanding of not only the needs, but also the contexts of use in which they emerge (Kujala 2008). To that effect, some qualitative studies have been published to account for users' thoughts and opinions when designing or trying to improve co-creative systems (Rezwana and Maher 2022b; Oh et al. 2018). But both studies focus on specific co-creative systems and do not tap into the deeper functional requirements of the users. Instead, they deduce how AI-to-human communication (Rezwana and Maher 2022b) or co-creative systems' interfaces (Oh et al. 2018) play an important part in human-AI collaboration and could be improved.

Finally, traditional qualitative studies such as the ones solely based on interviews are not perfect for RG for creative software due to the elusive nature of creativity. One technique that accounts for the unpredictability of relevant events, specifically designed with artists in mind, are cultural probes (Gaver, Dunne, and Pacenti 1999). A cultural probe consists of a physical package with all sorts of tools (e.g. notebook, camera, map) that artists can use freely to capture their moments of inspiration and ideas over an extended amount of time. Nowadays, phones (Bainbridge, Novak, and Cunningham 2010) and even wearables (Lin and Windasari 2019) can act as virtual cultural probes to study, for example, how their continued use can affect users' wellbeing. Still, to maximise the potential of the data gathered via a cultural probe, these are usually paired with other types of qualitative techniques such as interviews.

Methodology

As previously mentioned and as defended by Kujala (2008), the benefits of including the user in the early steps of development through qualitative fieldwork are well-known in the field of RE regarding software usefulness, usability, and acceptance. Accordingly, we propose to follow a cultural probe methodological approach complemented by two interviews to overcome some of the challenges in collecting and understanding general needs for creative software.

The first of these challenges, already mentioned, is related to the fact that the creative process is a continuous, unpredictable, and lengthy process that is hard to confine into a time-bound session. Additionally, it is important to deal with various types of media when attempting to describe general domain-agnostic needs. We also argue that using a cultural probe approach requires fewer resources, as it does not necessarily involve any prototypes or finished products. This in turn leads to more general observations, as the user is not limited to a specific software concept.

We now describe the two main phases of our methodology – data collection and data analysis – in more detail .

Data collection

A total of 21 people participated in our experiment, with ages ranging from 24 to 50 years old. From those, 12 identified as male, 8 as female and 1 as non-binary. In terms of creative fields, there were 2 people from Cinema, 1 from Cooking, 6 from Creative Writing, 3 from Design, 9 from Music, 3 from Painting, 1 from Photography, and 2 from Theatre. These fields overlapped for some people and 4 of them were also teachers. Participants had different levels of relationship with their art, ranging from fully professional to hobbyists. All data were gathered in May 2022.

We first conducted a semi-structured introductory interview (30-60min) with each participant, asking general questions about their education, occupation, and their views on creativity and creative activities. In this interview, we introduced the concept of our digital cultural probe, which we named POCket Artist (POCA), and that posed as a hypothetical creative machine capable of attending to whatever request was sent by the artist. Participants interacted with POCA through WhatsApp: we set up a dedicated phone number to which the artists could send messages. They were free to use any of the media formats currently available on the platform (audio, text, images and videos). There was only one guideline for the type of requests sent to POCA which was that it could only use the same means provided by the instant messaging application to answer back, e.g. if a participant were to ask for a cake, POCA could only send back a recipe for it and not the cake itself.

To avoid any frustration, the participants were explicitly informed that POCA would not execute their requests. In any case, they were instructed to send their wishes as if the program was able to provide the wanted result just the same and to think about it as their “ideal creative partner”. The upside to this is that participants were only limited by their imagination - and the platform’s interface as previously mentioned - regarding their requests, something we highlighted during the first set of interviews[1]. With this experiment, we expected to collect which types of necessities artists have during their day-to-day lives regarding creative activities and also which elements of their creative process they are willing to share with a machine.

After one week of interaction with POCA, another semi-structured interview was conducted (30-90min) to understand how they integrated the system into their routines and the hindrances to such integration. We also asked some questions about how the interaction with POCA affected their overall creative experiences, as well as some general questions about each participant’s perceptions of creativity, creative processes, and creative machines. Through these final questions, we aimed to consolidate the possibly unreported needs they felt during their week of interaction with POCA as well as to better comprehend artists’ worries and misconceptions regarding the field of CC. In the end, we presented a text2image tool, DALL-E 2 (Ramesh et al. 2022), and followed a use case exercise to understand more concretely how such generative tools could help.

In total, we collected 31 hours and 31 minutes of audio from the interviews, besides 135 interactions sent directly to POCA through WhatsApp in a wide range of formats: text, audio, images and photos, or even links to pages and videos.

Data Analysis

For data analysis, we followed a method for Thematic Analysis known as Provisional Coding (Saldaña 2021, p. 144), in which the data are categorised according to a predefined set of codes that can be obtained in many ways; in our study, we decided to use a pilot study for this purpose. We utilised Whisper (Radford et al. 2022) to automatically transcribe the interviews[2] and Notion[3] to facilitate the coding process.

Pilot Study For our pilot study, we first conducted an Open Coding process, segmenting the data from interviews and interactions with POCA for three participants into Units of Meaning (UoMs), each of which was then summarised into one to three sentences. Through Axial Coding, we drew connections between each of such descriptions according to their meaning and relation to the main research question, resulting in clusters known as categories. This process was done iteratively three times until we reached nine categories organised and described in depth in a codebook[4], a document containing a detailed description, inclusion criteria, exclusion criteria, points of confusion, and examples for each category, which would be used to guide the next step of the Provisional Coding. Each UoM could be included in more than one category (Saldaña 2021, p.80). The Open Coding process was performed by three researchers separately, but the entire Axial Coding was done collaboratively, to incorporate different aspects and views into the analysis.

Final coding Following the next step, the data of another 12 participants[5] were segmented into UoMs according to

[1]The scripts followed during the interviews are available at: `https://github.com/Superar/POCA`.

[2]Quotes were translated to English by the authors.

[3]Available at: `https://www.notion.so/`

[4]The codebook obtained during the pilot study is available at: `https://github.com/Superar/POCA`.

[5]Due to time and resource restrictions, we were not able to fully evaluate all data from the 21 participants. A total of 15 (3+12)

their relevance to our research question; we highlight that the material of a participant was not segmented by the researcher who interviewed them. Then, the researcher who had not yet seen the participant's data coded these UoMs according to the codebook, selecting the level of confidence they had in their decision (easy, hard, or very hard). Subsequently, we carried out the same Open Coding and Axial Coding process described above for the cases considered hard and very hard; this process resulted in a new organisation of the knowledge which required a re-coding of all data according to this new schema, culminating in the final themes we present in the next section of this paper: 12 themes representing user needs, 8 other regarding contexts of use, and 3 related to other aspects.

Findings and Discussion

Throughout the last axial coding, three different classes of themes emerged. Two of these classes directly reflect the theoretical separation between need and context of use presented by Kujala (2008) for which we present the respective themes in the next two subsections. In a single UoM, sometimes only the need might be addressed, leaving out the context (e.g. "Recommend a song"), other times the opposite might happen (e.g. "I need help with school"), but they are often together, even if implicit. In addition, themes in qualitative studies are not always clear and obvious. A natural consequence of this is that UoMs are usually multi-theme.

The last class of themes relates to several user comments on other non-functional components or even their general views on technology such as specific interaction requirements, use cases where they prefer not to use software, and even opinions on the applied methodology. These are exposed and discussed in the last subsection.

User needs

The first facet expressed by the themes describes the needs that users have related to the role the machine plays or is expected to play. For this purpose, we used the Kujala's (2008) definition of needs: "[...] problems that hinder users in achieving their goals in a specified context of use."

In total, we found 12 themes related to general user needs, each one corresponding to a specific role, that we summarized in Table 1 and proceed to elaborate.

Recorder A simple but prevalent need participants expressed was to have a recorder; a sort of digital vault to store mostly ideas, but also references and other types of information for later consultation. Usually, this recording was an end in itself - "POCA for me served as a place, a site, a method for a repository of ideas and relationships that I came across last week while researching or having some kind of idea connected to the creative process." (P4; Cinema); but sometimes this also happened when artists wanted to increase their productivity or were developing an artefact as well. Additionally, some participants felt the need to ask questions they did not want to share with other people or simply to vent: "[POCA] is just a vent, it will just be for me to vent what I need." (P18; Theatre, Creative Writing)

participants were analysed, corresponding to 416 UoMs gathered from 102 interactions and about 18h of recorded interviews.

Gatherer Artists mentioned, in many occasions, the necessity of finding already existing pieces of work or general information, e.g. product prices, term definitions, or study techniques. When asking for recommendations, this search is generally guided by some input: other artefacts (e.g. images, poems, songs), a concept – "[...] the concept of ghost time travel was the kind of thing I'd like to explore with POCA" (P4; Cinema) – or a piece of information (e.g. an author's name or a music genre). Users also ask for specific building blocks for their own work, such as words, colour palettes, fonts, excerpts of text or video, and templates.

Operator Machines are often required to perform the role of technical tools, producing reliable and predictable effects depending on user actions and parameters. Some examples are software for manipulating, composing, and editing content as well as software capable of creating simple products or stimuli, such as transcribers and metronomes. This also includes hardware interfaces, such as drawing tablets and Virtual Reality (VR) devices. As mentioned by a participant, these tools are already essential in their creative process: "[Software for manipulating scores and mixing,] [...] not being part of the final artistic product, [...] were already important to get to the final artistic product." (P21; Music, Teaching)

Generator Autonomous generation was an often sought need as well and appeared in all contexts we identified. Sometimes participants wanted to materialise some idea they had, develop some artefact or have a new source of inspiration. Participants also found it could be interesting to make up for their shortcomings outside their creative domain: "People like me, who can't [really] draw, often have this feeling that if they could draw, they would draw great things. [...] And [...] throwing ideas around and seeing them turn into images without knowing how to draw is a very cool thing, isn't it?" (P18; Creative Writing; Theatre) But other times participants wanted to generate different possibilities or solutions for chores they also have – like producing images for digital marketing – or even to help them with their personal lives – like generating recipes given a set of ingredients. Finally, many participants were unfamiliar with DALL-E 2 (Ramesh et al. 2022) but found its capabilities interesting, and even came up with new ideas for similar applications: "I [could give DALL-E 2] music. Because a lot of the time I make up stories while I'm making up a song [as an exercise for kids in my classes]." (P21; Music, Teaching)

Variator Presenting possibilities is a popular request, especially when the user already has a provisional but not perfect solution. In those cases, users seem to value the ability to provide several different and unexpected variations of one artefact provided as input, while keeping its most important features. This idea can take the form of a synonyms suggester, a prompt-guided image modifier, a paraphrasing tool, or even a stage direction planner: "it would be very handy, for example, to be given proposals of alternative [stage] routes for the same space" – (P18; Creative

Table 1: Summary of the themes regarding user needs

Name	Description
Recorder	User inputs something to be stored but expects no special output
Gatherer	User prompts for some possibly aggregated information that is dispersed in a repository
Operator	User provides controlled instructions that have a predictable effect on an artefact
Generator	User provides a prompt and expects a novel previously unexisting artefact
Variator	User provides a base artefact and an optional prompt, and expects alternative variants of the artefact
Mapper	User provides an artefact and expects an artefact in a different media
Completer	User provides an uncompleted artefact and expects extended or completed versions of the artefact
Analyser	User provides an artefact and expects an objective analysis or description of the artefact
Critic	User provides an artefact and expects a subjective opinion on the artefact
Instigator	User does not expect to interact, they expect the software to actively remember them
Organiser	User does not expect any input nor output but expects some background rearranging behaviour
Enabler	User has a problem and expects to find solutions through continuous brainstorming interaction

Writing, Theatre). These alternatives are always evaluated and filtered by the user to be either used directly or to inspire a new user-made variation that is more adapted to the new use case or to the new constraints.

Mapper The task of translating an artefact into a whole new media is not usually straightforward, due to the several ways you can encode characteristics of the original artefact into the new one. Yet, participants reported that generating songs from images, images from poems or "for example, to provide a song and receive an image would be something very interesting." (P4; Cinema) The participants want the outputs to complement the original artefact, for example in social media, or even to further inspire them by contrasting with previously developed ideas based on the original artefact. There is also great applicability of mapping tools to make art more accessible. By making an artefact available in several media the artists depict a reality where they could effortlessly reach a wider public and allow some people that could not appreciate their artefacts before due to sensorial limitations to finally experience .

Completer Participants also refer the need of completing artefacts. It might be related to a lack of expertise, motivation or time to do a specific part, or even because using other generated components may be crucial or beneficial. One of the mentioned cases was related to the specific surrealistic collaborative method, the exquisite corpse: "I like to create alone but that way I can't do exquisite corpses. Besides, scheduling sessions [with people] is hard. I want some entity that is always available for artistic partnerships. Can you complete my drawing? [...] Or my text? Or my song?" (P3; Painting) Tasks like image inpainting, extending or completing a musical opus, or finishing a rhyme scheme are great examples of use cases. Sometimes artists want the system to perfectly mimic the provided part, other times they ask for contrasting elements or even a mix of both. Lastly, users prefer to use the system's output as it is, although sometimes participants might want to decide upon its quality possibly asking for new completions or variations of existing ones.

Analyser Another task users need the machine to perform is analysing a given piece of work, extracting and making some of its intrinsic characteristics explicit. An example is textual analysis – "If [POCA] could go through the whole document and identify precisely those [word] repetitions, so I could correct them later, then that would be great!" (P7; Creative Writting) Other examples are the identification of shapes and lines in paintings or the recognition of music tempo and progression. There were cases in which the artist wants the machine to analyse themselves by understanding their behaviour (e.g. when they are more productive) or by extracting characteristics of a performance (e.g. movements made while conducting an orchestra).

Critic Having a personal critic in some sense that could provide feedback or a second opinion was also an observed need. It is closely intertwined with the role of Analyser, where the user inputs an artefact, but differs in the sense that the expected output here is an opinion. For example: "I like to receive input, not necessarily from a human being. Just the fact that [POCA] could give some kind of feedback of ideas, of concepts [...] could be enough to make what I do creatively a little richer." (P3; Painting); "[...] what would help me a lot [...] is to have proofreaders who go a little beyond identifying typos [...], saying if the text is well constructed, if in that language it makes sense [...], to be my little pocket proofreader." (P7; Creative Writing)

Instigator A majority of creators mention a program that could challenge or remind them to be creative, for example, by imposing conditions on their creative process or by providing daily/weekly challenges: "I think it would be very interesting to have a program that you open and it throws you into a totally different situation than the one you are in now. In a way that makes you uneasy enough to be able to produce thoughts and emotions." (P19; Painting, Design); "Maybe a machine that would help me do [a creative writing] exercise every day would be a way to stimulate my creativity for writing." (P12; Music, Creative Writing). The machine should also actively remind artists of important tasks and relevant ideas in an appropriate time, which can help them feel more motivated to extra develop their creativity.

Organiser Keeping things organised is always a time-consuming task: "I have difficulties in organising [...] [so] all my requests [to POCA] had to do with that, a way to help me organise my ideas and bringing them back another time." (P2; Painting) Even when there are ways to quickly record things, that does not mean they will be organised, and many times entail a posterior organisation process, which can be overwhelming. The act of recording is not the relevant part of this scenario, nor is the moment when the user wants to gather some of that information back. Instead, this role focuses on the active processes the users think the system should undergo to self-organise a repository according to its user's goals, by actively finding relationships among the data, identifying collisions, possible problems, missing information or ambiguous constraints. This involves crawling in other repositories to rectify or complement the available data, or even calculating metrics for the user; and includes use cases such as organising finances, projects, schedules, documents and references, playlists, or even a shopping list. Such different use cases make these processes quite complex to adapt to each user's necessity. Besides, the organisation process sometimes cannot be externalised to a virtual assistant or secretary as it is a vital part of the creative process.

Enabler During the experiment, artists sometimes felt stuck, without knowing what to do next, so they expressed wanting help with this by discussing their artefacts, brainstorming ideas, or being provided references for inspiration. In essence, they wanted help enabling their creative process. For example, one participant said "[...] during the execution of this work, [which is] already in progress. At some point I would stop and think that I don't really know how to proceed with it. So I went and brought to POCA how I can move forward with that. Not necessarily expecting obvious answers" (P2; Painting). Finally, this need emerged in all but one of the contexts of use, including wanting help unlocking personal life decisions, improving personal development, or solving logistics tasks and problems.

Contexts of use

Recalling the definition by (Kujala 2008), needs arise in specific contexts of use. In this subsection, we strip the former from the UoMs and focus on the latter. While a need is something actionable, a context is defined by the surrounding circumstances and is often relative. The same user need can appear in two distinct contexts, for example, a generated picture can inspire the user or be shown in a class.

Overall, eight different contexts of use were gathered from the data. We now provide further details on them.

Inspiration Inspiration is often seen as the spark that ignites the creative process, providing the artist or creator with the initial idea or vision for their work. So it is only natural this was one of the most mentioned contexts, especially when gathering references and information to inspire artists: "So I go to museums [related] to the themes that I want to work on at the moment. [...] And I get a lot of visual references. And after that I manage to create some things." (P19; Painting, Design) Other times, artists simply wanted to record a reference they found inspiring or sought inspiration by asking to be challenged or stimulated, for example by autonomous generation: "Today, I would like to illustrate stories created [from keywords]. If I gave the elements 'elephant', 'tea' and 'vampire', what would [POCA] give me to illustrate?" (P3; Painting) Finally, sometimes participants did not want just mere references, but to be fully immersed in a different reality or ambience that allowed them to be fully inspired by it: "So we want this play to be a play with medieval characteristics, okay? [POCA] will look for the medieval historical context, [...] the type of clothing [...], [...] the type of music [...] and bring already these tools to study and to create [...]" (P18; Creative Writing, Theatre)

Ideas If inspiration is the spark that ignites the creative process, ideas are its fuel. They might be hard to find, or even be already present subconsciously. The biggest need related to ideas is having a way to store them somewhere so they do not get lost in the future, and ideally they would be organised automatically. The flip side to this is the need to be able to access these ideas effortlessly and at any time, from one's portfolio. Other needs related to ideas are discussing, evolving, bringing about, reminding, instigating, and exploring ideas in collaboration with creative software to unlock the creative process: "Sometimes I thought, [...] if I take this idea and try to explore in POCA, [...] it gave me [...] that freedom. [...] I think POCA can also serve as an option for [...] exploration." (P13; Music)

Artefact Development When people think about artists' needs, usually what comes to mind is the materialisation of creative ideas into the final piece of work – which can be material, performative or any other format. A machine can perform various roles in this context, from helping the artist to organise their process and drafts – "[POCA] would enable me not to get too disorganised and to know what I still had missing" (P18; Creative Writing, Theatre) – to allowing them to overcome their limitations – "I am a lyricist-composer. So, my greatest aptitude is writing the lyrics. But not the harmony. So, a program that would give me the harmonic options [...] Gee, that would help me a lot." (P12; Music, Creative Writing) Artists already use technology to produce and manipulate their work and find all kinds of information; but they still also seek alternative ways of thinking and collaborating. Conversely, some artists fear that the human aspect of art might get lost due to technology: "I wouldn't have experienced nor gained a shred of what I experienced with the person who made those drawings. [...] That accompanied a human process between us, seeing another person feeling what I wrote and transforming it into drawings through their own sensation." (P18; Creative Writing, Theatre)

Productivity Another prominent context of use is to manage (access, organise, save, or control) some kind of resource, e.g. material, people, venues, time, motivation, or energy. Artists mention explicitly that they could use computational tools in such context to concentrate on their creative tasks and be more productive: "Something that would

[...] not let you procrastinate for so long. Or prioritise those creative ideas a bit more, that would be [...] too exciting." (P5; Music, Creative Writing) The machine can help with unavoidable adjacent tasks, usually not directly related to the artist's own main creative process, such as bureaucratic matters or even "with the creative tasks which I take no pleasure in and have to do anyway" (P21; Music, Teaching).

Skill Development Expertise, practice and skills are crucial parts of being an artist, and teachers play a crucial role in developing those. Maybe, that is the reason why participants would like software to fulfil that role: "just like a real teacher, who I could contact, anytime, to ask something, make something clear, evaluate what I did, provide a reference." (P19; Painting, Design) In this context, the software can also detect difficulties, provide examples and feedback, manage study time, provide methods or even suggest exercises. These exercises can be directed for a specific hassle, technique, standard, domain, or even intended to stimulate general creativity or other mental capacities. Two non-exclusive kinds of general creativity exercises are referred to: those that impose constraints on the process; and those that provide inspiring stimuli or concepts as a starting point. These exercises not only can make the artist learn how to deal with creative blocks, but also learn more about themselves and their limitations: "one thing that would help me [...] would be the identification of moments of optimal creativity. [...] What I should do to become more inspired and/or produce more." (P3; Painting)

Teaching Artists, while being eternal students, are also experts and consequently often turn out to be teachers as well. Teaching has its own challenges, and software can play a big role in making it easier to handle. Participants refer to remote classes and reference gatherings as cases where they already use software. Other new cases where software can become handy are the emulation of conditions for an evaluation or specific performance, the provision of personalised feedback for students or the analysis of how teachers can convey their ideas more effectively. Generation of content for classes is yet another great request: producing slides, notes or summaries, creating more fun and motivational activities for children, or even customised exercises: "for example, I need to create some sheet music reading exercises for this student, [...] a simple reading, [...] I can simply go search in a book [...] but I always think: let's try to do something more customised to fit her difficulties. So I could have an assistant to which I could say: "[...] my student has these difficulties [...] I need two reading exercises ", [...] and it generates them." (P5; Music, Creative Writing)

Social interaction Technology is a powerful tool to foster social interaction and cooperation – "[...] this tool could be used from the various perspectives of the various stakeholders [...] and they would basically be working on a common platform, simultaneously, for the same thing." (P18; Creative Writing, Theatre) On the other hand, there is also a concern that digital communication is harmful to general social interaction: "This amputation of the human and communication capacities with each other that we are voluntarily doing [...] is deeply harmful [...] because these people have no means of expressing themselves." (P18; Creative Writing, Theatre) Sometimes the user does not want to interact with other people and the machine might be a way to avoid that – "I had a place where I could ask all the questions without holding back, without upsetting anyone. And I could be inside that question as long as I wanted and not as long as the other person wants." (P21; Music, Teaching)

Personal Life Many participants refer that the best way software can help in their creative tasks is to help them have more time for the creative process, by taking over other necessary tasks inescapable to humans: "if it could do those things, to help me have more time, to optimise my time [...] We have time, right? But we have so many things to do, that if it could do those small things for us [such as groceries list], that would be great." (P12; Music, Creative Writing) This includes deciding what to buy, where to buy it for the best price, how to get there, and what and how to cook with it afterwards. Other tasks can include helping users deal with their insecurities and negative feelings such as tediousness, frustration and anger, providing company while sharing their positive emotions, or motivating and encouraging them. In short, participants refer that software seems potentially useful in making the user feel well and comfortable in general, by recommending new habits or customised new experiences (such as a new haircut style or new music to hear) while also allowing the user to have the last call.

Other Aspects

Interaction When talking about the functional requirements with participants, aspects pertaining to how users would prefer to interact with creative software inevitably arose. Since there is much more available literature on this topic than regarding functional aspects and since this was not our main focus, we decided to keep the study of these aspects to the minimum necessary to explain the benefits of our methodology. However, we believe they deserve a longer and deeper analysis. Nonetheless, it is relevant to notice that despite the limited interaction model provided by WhatsApp that we instructed in the first interview, participants were still able to devise new and different modes of interaction such as real-time interfaces or even a VR partner.

There were several different aspects that were mentioned by the participants regarding non-functional requirements for their ideal creative tool : autonomy, intrusion, involvement, availability, adaptability, customisation, expertise, learnability, human likeliness, social and emotional skills, collaboration, cooperation... In short, participants agreed that an ideal POCA should be available, adaptable and easy to use, as human-like as possible, while also allowing users to overcome natural or communication barriers without ever subordinating human-to-human social interaction.

Non-Necessity It is important to consider in which situations and why some participants in the study did not feel the need to use POCA. This is a significant aspect of our research, as it contradicts the assumptions that machines should and will be used by artists and bodes well for Gasson's (2003) argument supporting human-centred design.

Our analysis revealed that some participants rejected the idea of human-like creative machines, especially when the machine is perceived as an alternative to human creativity, personal experiences, and feelings – "[...] computers are a human creation [...], but they should be at the service of the human, it should not be an alternative to the human." (P18; Creative Writing, Theatre) Participants also did not want to substitute their own creative process with POCA: "[...] if I'm substituting myself for another mechanism, I'm taking away some of the pleasure that gives me that process, isn't it?" (P21; Music, Teaching) Another aspect to highlight are the limitations the artists assumed the machine had: "I think I was left thinking a lot about what would be doable or not" (P2; Painting); "it is not possible to be bold because there is a technological limit to boldness." (P21; Music, Teaching)

On the other hand, some participants viewed technology as an opportunity to adapt: "we have to move around, we have to be artists in other ways, just as many others have been, the contexts change" (P19; Painting, Design). Additionally, machines present an important aspect of democratisation, allowing the artist to overcome technical and accessibility barriers. While they prefer to not replace their own creative process with software, they do not like depending on others who might not want to outsource their creativity. For example: "[...] instead of having a person with a computer in hands and feeling that I was really being heavy, [...] there could be some system here that allowed me to walk by her side without carrying the weight [...]. Depending on others is an anguishing thing" (P18; Creative Writing, Theatre)

Methodology The single fact that we were able to gather evidence about when users prefer not to use software and where people detached the imposed interaction limitations, re-imagining other kinds of interaction, as explored in the last two subsections, is in itself already a sign of the benefits of this methodology. However other benefits and limitations were also explicitly mentioned.

Qualitative studies usually have a direct impact on how users perceive their tasks and methods. Our study was no different: artists felt the study itself had a beneficial impact on the perception of their creative process and the role of technology in it, often even motivating them to exercise more of their creativity. Having a familiar, available and accessible cultural probe which can manage several media was also explicitly pointed out as an advantage: "The idea of bringing this to WhatsApp, for example, is phenomenal. Basically, everyone that I know, artist or not, uses it." (P15; Creative Writing) "My grandma with 80 plus years and being illiterate uses WhatsApp." (P15; Creative Writing)

On the other hand, some still mentioned that WhatsApp is not the best tool to precisely and effectively discuss certain artefacts. Besides, and despite the alerts, people still felt underwhelmed when interacting with POCA, since it could not fulfil the provided demands, leading participants to forget about it throughout the week. Some solutions to this problem would be using a functional prototype, openly available software, or even a person. However, all these might either imply additional costs or confine the user to focus on the interaction instead of the potentially available functions. Related to this point, some participants mentioned that they were not able to abstract from what they thought a machine could do: "POCA is so abstract that is not so easy to understand in a concrete form." (P18; Creative Writing, Theatre) Besides forgetting about it, people also reported moments when they did not interact with POCA because they did not want to interrupt their creative immersion, felt embarrassed, or thought a machine was not capable enough. Yet, these moments of no interactions still allowed the user to ponder on those needs and allowed us to discuss those cases on the last interview. The last limitation of our approach was the duration of the collection: "one week [...] is too short." (P18; Creative Writing, Theatre)

Conclusion

In this paper, we bring forward a methodology that we used to uncover a set of 12 general user needs and 8 contexts of use, which allowed us to better understand what artists want from their current and future creative tools. We consider both the proposed methodology and the user-informed findings to be a step forward in involving artists in the design of creative software and the field of CC.

This methodology can be adapted and applied with an existing specific system in mind or even in the early stage of development of a new (co-)creative system, by confining it to one creative field, for example. In this case, new and more specific needs or contexts might be uncovered that can be used to specify tailored and detailed domain-specific functional requirements. In any case, our own themes already provide an ample base to classify functional requirements. They can also be used as a system categorisation framework, assessing which needs and contexts are addressed by a system, in this way better highlighting its strengths and exposing its shortcomings. For example, ChatGPT (Ouyang et al. 2022) can generate new text fairly easily and reliably (Generator), but it cannot really gather resources or sources of information consistently or trustworthily (Gatherer). For systems that have yet to be built, our themes are also useful to serve as a starting point for design exploration.

Finally, further research includes the analysis of the remaining data that could bring forth new information, although we believe we achieved data saturation with 15 participants. Another direction can be the exploration of some of the limitations identified, for example, comparing our results with results using a responsive POCA, even if by means of a Wizard of Oz (WOZ) (Thelle and Fiebrink 2022). Another evident direction would be to extend the duration of the data collection. We also restricted this study to the topic of artists' needs, but the data contained information on other aspects such as interaction features, or even artists' emotions, perceptions, and fears, all of which are valid and pertinent directions to follow too. Ultimately, a more formal definition of our themes and framework implementation could be advantageous, as well as comparing our themes with current theoretical interaction frameworks for co-creativity (Kantosalo et al. 2020; Rezwana and Maher 2022a; Kantosalo and Jordanous 2021) and studying how they can be used jointly.

Author Contributions

All authors contributed to the concept and execution of the study, including contacting the participants, carrying out interviews, segmenting, and analysing the data. This paper was also written and revised by every author.

Acknowledgments

We would like to express our gratitude to Professor Licínio Roque for motivating and counselling this research, our colleagues from the Research Methods class for their useful feedback, and our supervisors, Professors Amílcar Cardoso, Mathew Davies, Hugo Oliveira, and Geraint Wiggins, for their guidance and support. We would also like to thank the participants of the study, whose invaluable contributions made this research possible. Their cooperation and insights helped us to produce a high-quality research output. This work is funded by national funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020. The authors are also individually funded by Foundation for Science and Technology (FCT), Portugal, under the PhD fellowship grants (2021.05529.BD, 2021.05653.BD, UI/BD/153496/2022).

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. 2023. MusicLM: Generating Music From Text.

Ahmad, K.; Bano, M.; Abdelrazek, M.; Arora, C.; and Grundy, J. 2021. What's up with Requirements Engineering for Artificial Intelligence Systems? In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, 1–12. Notre Dame, IN, United States: IEEE.

André, P.; schraefel, m.; Teevan, J.; and Dumais, S. T. 2009. Discovery is never by chance: Designing for (un)serendipity. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition*, 305–314. New York, NY, United States: Association for Computing Machinery.

Bainbridge, D.; Novak, B. J.; and Cunningham, S. J. 2010. A user-centered design of a personal digital library for music exploration. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, 149–158. New York, NY, United States: Association for Computing Machinery.

Bano, M., and Zowghi, D. 2013. Users' involvement in requirements engineering and system success. In *2013 3rd International Workshop on Empirical Requirements Engineering (EmpiRE)*, 24–31. Rio de Janeiro, RJ, Brazil: IEEE.

Colton, S.; Pease, A.; Corneli, J.; Cook, M.; Hepworth, R.; and Ventura, D. 2015. Stakeholder Groups in Computational Creativity Research and Practice. In Besold, T. R.; Schorlemmer, M.; and Smaill, A., eds., *Computational Creativity Research: Towards Creative Machines*, volume 7. Paris, France: Atlantis Press. 3–36.

Gasson, S. 2003. Human-Centered vs. User-Centered Approaches to Information System Design. *Journal of Information Technology Theory and Application (JITTA)* 5(2):29–46.

Gaver, B.; Dunne, T.; and Pacenti, E. 1999. Design: Cultural probes. *Interactions* 6(1):21–29.

Heyn, H.-M.; Knauss, E.; Muhammad, A. P.; Eriksson, O.; Linder, J.; Subbiah, P.; Pradhan, S. K.; and Tungal, S. 2021. Requirement Engineering Challenges for AI-intense Systems Development. In *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, 89–96. Madrid, Spain: IEEE.

Kantosalo, A., and Jordanous, A. 2021. Role-Based Perceptions of Computer Participants in Human-Computer Co-Creativity. In *7th Computational Creativity Symposium at AISB 2021*, 20–26. London, United Kingdom: AISB.

Kantosalo, A.; Ravikumar, P. T.; Grace, K.; and Takala, T. 2020. Modalities, Styles and Strategies: An Interaction Framework for Human–Computer Co-Creativity. In *Proceedings of the 11th International Conference on Computational Creativity*. Coimbra, Portugal: Association for Computational Creativity.

Kujala, S. 2008. Effective user involvement in product development by improving the analysis of user needs. *Behaviour & IT* 27:457–473.

Lin, F., and Windasari, N. A. 2019. Continued use of wearables for wellbeing with a cultural probe. *The Service Industries Journal* 39(15-16):1140–1166.

Oh, C.; Song, J.; Choi, J.; Kim, S.; Lee, S.; and Suh, B. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. Montreal, QC, Canada: Association for Computing Machinery.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 27730–27744.

Pei, Z.; Liu, L.; Wang, C.; and Wang, J. 2022. Requirements Engineering for Machine Learning: A Review and Reflection. In *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, 166–175. Melbourne, Australia: IEEE.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents.

Rezwana, J., and Maher, M. L. 2022a. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction* 27.

Rezwana, J., and Maher, M. L. 2022b. Understanding User Perceptions, Collaborative Experience and User Engagement in Different Human-AI Interaction Designs for Co-Creative Systems. In *Creativity and Cognition*, 38–48. Venice, Italy: Association for Computing Machinery.

Rodriguez, G.; Wong, L.; and Mauricio, D. 2017. A systematic literature review about software requirements elicitation. *Journal of Engineering Science and Technology* 12:296 – 317.

Saldaña, J. 2021. *The Coding Manual for Qualitative Researchers*. Los Angeles, CA, United States: SAGE Publications Inc., fourth edition.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; and Taigman, Y. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *Eleventh International Conference on Learning Representations*. Kigali, Rwanda: OpenReview.

Thelle, N. J. W., and Fiebrink, R. 2022. How do Musicians Experience Jamming with a Co-Creative "AI"? In *Workshop on Machine Learning for Creativity and Design (NeurIPS 2022)*.

Wirtz, R., and Heisel, M. 2019. RE4DIST: Model-based Elicitation of Functional Requirements for Distributed Systems. In *Proceedings of the 14th International Conference on Software Technologies*, 71–81. Prague, Czechia: SCITEPRESS - Science and Technology Publications.

Differentiable Quality-Diversity for Co-Creative Sketching AI

Francisco Ibarrola and **Kazjon Grace**
School of Architecture, Design and Planning
The University of Sydney
Sydney, Australia
[francisco.ibarrola,kazjon.grace]@sydney.edu.au

Abstract

Co-creative artificial intelligence systems go beyond one-click generative AI solutions and enable users to participate in the generative process. A key component of co-creative interfaces is the ability to suggest multiple options to the user, to avoid constraining the process and help overcome creative blocks. We explore this diversification problem in a vector drawing synthesis-by-optimisation setting and propose algorithms for generating diversity among user-defined characteristics. Experimental results show improvement in terms of behaviour coverage and image diversity.

Introduction

Co-creative systems built on generative AI models expand upon the latter's increasingly impressive expressive capacity by offering their users additional control and creative agency. This additional agency is critical in the early stages of creative tasks across a wide variety of domains. In creative professions such as art, interface design, engineering, and architecture, sketching is an important aspect of this early conceptual exploration (Goldschmidt 1991; Gero 1998). At the outset of the creative process, artists and designers typically lack a clear idea of what they are looking for, or accept that their current ideas may change entirely as the process progresses. In the context of design, those who approach creative tasks without this level of ideational flexibility often fail to achieve their goals (Dorst 2015), while those who embrace flexbility have been shown to produce more-creative output (Suwa, Gero, and Purcell 2000). This is because designing, especially in its early stages, is not a mere process of synthesising given requirements, but rather an iterative process of discovering and refining both those requirements and how they might be fulfilled. This has been described as a co-evolution: a concurrent emergence of both the problem (the requirements) and its solution (the design) (Poon and Maher 1997). Current research in co-creative systems is exploring how this co-evolution can be supported by AI tools (Lawton et al. 2023; Gero, Liu, and Chilton 2022; Williford et al. 2023).

While the notion of "diversity" depends on context, this paper operationalises co-creative diversity with reference to some observable characteristics quantifiable by numerical values, that can be chosen depending on the context. The values of any given output can then be thought of as a multidimensional point in what is usually called a behaviour space. Subsequently, the breadth of the distribution of a set of points in the behaviour space provides a domain-appropriate measure of diversity. This is related to the notion of a generator's expressive range (Smith and Whitehead 2010), although there user-chosen characteristics are used to assess a generator's diversity in a space, while the term "behaviour space" derives from the quality-diversity (QD) literature, where generators actively seek to cover their behaviour spaces (Pugh, Soros, and Stanley 2016).

In this paper we introduce this QD approach to the CICADA model (Ibarrola, Lawton, and Grace 2022), a drawing agent designed to work cooperatively with a human designer. Previous analysis of user experience with a CICADA-based co-creative system suggests that the capacity to select from between different drawing options would be of value to users (Grace, Lawton, and Ibarrola 2023). Given that CICADA consists of an end-to-end differentiable generation-by-optimisation process, we build on previous approaches for enforcing diversity in a differentiable setting, such as OMG-MEGA (Fontaine and Nikolaidis 2021). This algorithm stochastically explores behaviour space, but we show it to be ill-suited for our context on account of the properties of CICADA's parameter space, and propose an alternative for this kind of settings. In this paper we describe OMG-MEGA along with its shortcomings in our context and propose two better-suited variants.

Method

The general setting of our approach consists of a parameter space Y (or the genome space in evolutionary computation parlance), associated to a differentiable generative model $h : Y \to X$. Additionally, two differentiable functions $f : X \to \mathbb{R}$ and $g : X \to B \subset \mathbb{R}^N$ map the elements of X into an objective score (or fitness) in $\mathbb{R}$ and behaviour dimensions $b \in B$, associated to some characteristics of $x \in X$ in which we want diversity.

We pursue the issue of finding diverse alternatives in a generative process as an exploration of the behaviour space B. That is, we want to generate a set of solutions that "behave" differently in terms of the outputs of g. We explore this problem in the setting of co-creative design using CICADA (Ibarrola, Lawton, and Grace 2022), where the pa-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

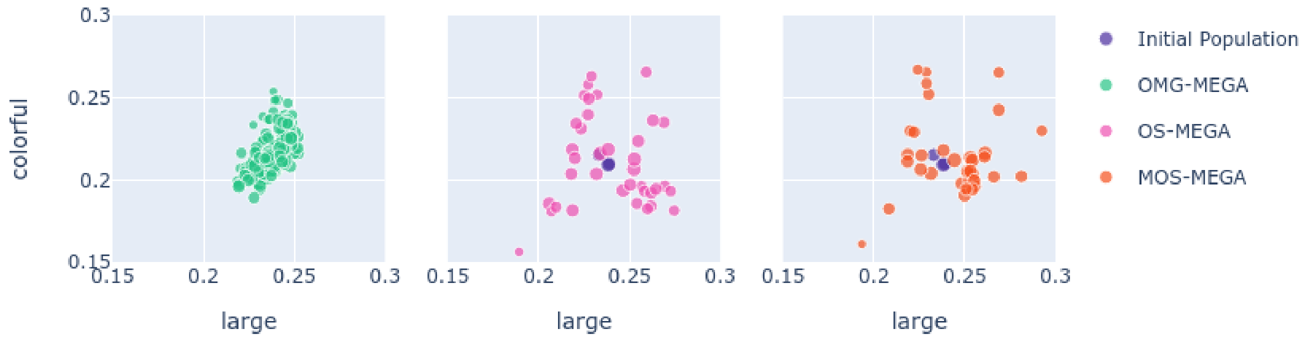


Figure 1: Distribution of populations generated with different algorithms in the (colorful, large) behaviour space.

rameter space $Y \subset \mathbb{R}^M$ is the set of arrays of parameters that determine the width, color and spatial locations of the set of Bézier curves that constitute a sketch or drawing. Also, h is a differentiable rasterizer (Li et al. 2020) that generates an RGB image x from an array of parameters y. Additionally, we define the objective function

$$f(x; t) \doteq \langle c_{img}(x), c_{txt}(t) \rangle,$$

as the normalised inner product between the CLIP (Radford et al. 2021) encoded latents of the image x and a text prompt t, provided by the user as the drawing's description.

In a similar fashion, we can define behaviours in terms of CLIP losses by addressing how well x matches additional characteristics. In our experiments below, we define

$$g(x) \doteq (\langle c_{img}(x), c_{txt}(\mathit{large}) \rangle \, , \, \langle c_{img}(x), c_{txt}(\mathit{colorful}) \rangle \,)$$

to set CLIP interpretations of size and chromatic variance as the behaviour dimensions.

We shall then consider the problem of producing a population of CICADA-drawn images $\{x_1, \ldots, x_K\} \subset X$ such that the elements x_k are maximally behaviourally diverse (i.e. in this case small-to-large and colourless-to-colourful).

OMG-MEGA

A recently developed and end-to-end differentiable approach to the quality-diversity problem is the Objective and Measure Gradient MAP-Elites via Gradient Arborescence (OMG-MEGA) algorithm (Fontaine and Nikolaidis 2021) ('measure' is the authors' term for what we call a behaviour). In brief, this consists of iteratively picking an existing element x_k from an existing population, generating a new individual by modifying y_k according to

$$y_{K+1} = y_k + |\alpha_0| \nabla_y f(y_k) + \sum_{n=1}^{2} \alpha_n \nabla_y g \circ h(y_k),$$

where $\alpha_n \sim \mathcal{N}(0, \sigma I), \forall n = 0, \ldots, N$. In other words, by using gradient descent over both the objective and the behaviours, but with random weights (positive-only in the case of the objective).

While this model is really good for thoroughly exploring the behaviour space, it is not very well suited for the CICADA problem, where drawings typically spend hundreds of iterations progressing towards recognisable shapes. We contend that for synthesis-by-optimisation tasks such as ours OMG-MEGA does not allow for significant enough per-iteration changes to each individual before randomly varying the objective and behaviour weightings, which jeopardises the algorithm's capacity to converge on images that are recognisable as a representation of the prompt.

OS-MEGA

To address the lack of convergence in OMG-MEGA on CICADA tasks, we propose to optimise for longer between selecting new coefficients. We also wish to avoid duplicating work by re-searching areas that have already been well-traversed, so we additionally want to enforce the random coefficients to be biased towards directions that purposefully lead them away from explored regions of behaviour space.

We start by picking an element at random from the current population, and build a new individual by moving away from the population centroid in the direction of least variance. That is, let $\{b_1, \ldots, b_K\}$ be the set of two-dimensional behaviour scores of the population, such that $g(x_k) = b_k \in \mathbb{R}^2$, and let $\bar{b}$ and C be the associated empirical mean and covariance matrix, respectively. Also, let v be the eigenvector of C with the smallest eigenvalue, i.e. the direction the population is least diverse in. Then, we can build a new individual by starting with $y = y_k$ and iteratively running

$$y' = y + \nabla_y f(y) - \lambda \nabla_y \| g \circ h(y) - b_k - \sigma v \|^2,$$

where $\lambda > 0$ and σ are weighting parameters, and the sign of σ is the sign of $\langle v, b_k - \bar{b} \rangle$, meaning the optimisation process is directed "outwards" from the explored area along the direction that has been least explored thus far. We refer as this algorithm as Outbound Scattering MEGA or OS-MEGA.

MOS-MEGA

In addition to directing the search in the behaviour space "outwards", the characteristics of our problem space suggest we may be able to improve on the diversity of the search by directly introducing noise through smart "mutation" strategies. OMG-MEGA replaces the traditional "mutation" genetic operator designed to introduce genetic diversity with the random coefficients on the objective and behaviours,

ISBN: 978-989-54160-5-9

which have the effect of adding noise to the search. However, the resulting changes are local and small-scale, and in our vector image context larger changes may be more effective. We take advantage of our vector representation to modify paths that are not significantly contributing to the objective, replacing the least-contributing with new, random traces. This operation is not differentiable, so we perform it after choosing an individual from the population but before conducting gradient descent.

Let us assume we have chosen an individual k, and let $P = \{p_1, \ldots, p_J\}$ be a partition of the set of parameters y_k, such that every p_j contains the parameters of a single trace. Then, we can compute a set of "irrelevance scores" $\{s_1, \ldots, s_J\}$ where s_j is the objective score of the image generated from y_k after subtracting (i.e. not drawing) the j-th trace. A low value of s_j means that discarding the j-th trace undermines the quality of the drawing. Consequently, we may improve diversity by adding Gaussian white noise with variance s_j to every p_j, obtaining a drawing which maintains the relevant traces of the original but differs in those that do not significantly contribute to the objective.

From here on, we can proceed with the gradient descent iterations as in the OS-MEGA algorithm. The full process (which we call Mutated Outbound Scattering MEGA or MOS-MEGA) is outlined in Algorithm 1.

Algorithm 1 MOS-MEGA

Initialization
Build starting population $\{x_1, \ldots, x_K\}$
$B = [g(x_1), \ldots, g(x_K)]$
for $t = 1, \ldots, T$
  Mutation Phase
  Choose a random $k \in \{1, \ldots, K\}$
  Compute the irrelevance scores $\{s_1, \ldots, s_J\}$
  for $j = 1, \ldots, J$
    $p_j \leftarrow p_j + \eta,\ \eta \sim \mathcal{N}(0, s_j^2)$
  $y = [p_1, \ldots, p_J]$
  Optimisation Phase
  $\bar{b} = \text{mean}(B)$
  $C = (B - \bar{b})^T (B - \bar{b})$
  $v =$ eigenvector with min eigenvalue of C
  $\sigma = \text{sign}(\langle v, b_k - \bar{b} \rangle)$
  for $i = 1, \ldots, I$
    $y \leftarrow y + \nabla_y f(y) - \lambda \nabla_y \| g \circ h(y) - b_k - \sigma v \|^2$
    $K \leftarrow K + 1$

Results

In order to test how the proposed algorithms work, we run a few examples using CICADA, starting from a partial sketch of "a red chair" and generating three random completions, whose behaviour scores are depicted in blue in Figure 1. From there, we run OMG-MEGA, OS-MEGA and MOS-MEGA. We run all algorithms for 1,000 seconds, to be able to fairly compare their performance, as they are intended to use in a co-creative setting, where time is a relevant factor.

The results are illustrated in Figure 1, where it can be seen that while OMG-MEGA produces more individuals, both of our proposed variants, OS-MEGA and MOS-MEGA cover a larger area of the behaviour space.

Figure 2 shows some (randomly chosen) examples of the actual images obtained, making it clear that the observed greater coverage in Fig. 1 translates to much more visible variance in the images. As previously stated, OMG-MEGA does not produce significant variations between CICADA images (at least not without prohibitive amounts of compute), whereas OS-MEGA significantly changes the characteristics of the drawings, and MOS-MEGA moreso again.

Visual comparison has a high degree of subjectivity, so we have made use of the Truncated Inception Entropy (TIE), as introduced in (Ibarrola, Lawton, and Grace 2022), to quantify the diversity of each of our resulting sets of drawings. TIE uses the same feature space as the well-known FID image quality measure, but assesses variance rather than comparing two sets of images. This is computed as

$$\text{TIE}(A; K) \doteq \frac{K}{2} \log(2\pi e) + \frac{1}{2} \sum_{k=1}^{K} \log \lambda_k, \tag{1}$$

where A is the population being evaluated, λ_k are the eigenvalues (in descending order) of their covariance matrix, after mapping with an inception network, and K is a truncation parameter. High values of this metric are associated with high population diversity.

Four completion tasks of sketches of common household items ("a chair", "a lamp", "a hat" and "a blue dress") were run starting with the same populations for the three algorithms, and the obtained TIE scores (using $K = 16$) are shown in Figure 3. The larger TIE values obtained with OS-MEGA, and larger still using MOS-MEGA corroborate the effects observed in Figure 2. Some examples of the obtained results can be seen on Figure 4, showcasing variety in size and chromaticity.

Conclusions

In this paper we have proposed new ways to explore behaviour space in the setting of co-creative drawing based on vector image optimisation. Experiments show that the proposed algorithms result in better coverage of behaviour space in the same CPU time, as well as greater diversity as attested by visual inspection and the TIE diversity metric. While our explorations have thus far focused only on the CICADA drawing context, we hope they may generalise to other quality-diversity contexts.

Future work will focus on how well the proposed algorithm works in a real co-creative setting, including both the time taken to usably generate different suggestions and the user ratings of their appropriateness and utility. Furthermore, additional studies are needed to explore the difficulties that may arise when the users are to define the behaviour space on their own (i.e. providing arbitrary prompts for both objective and behaviours).

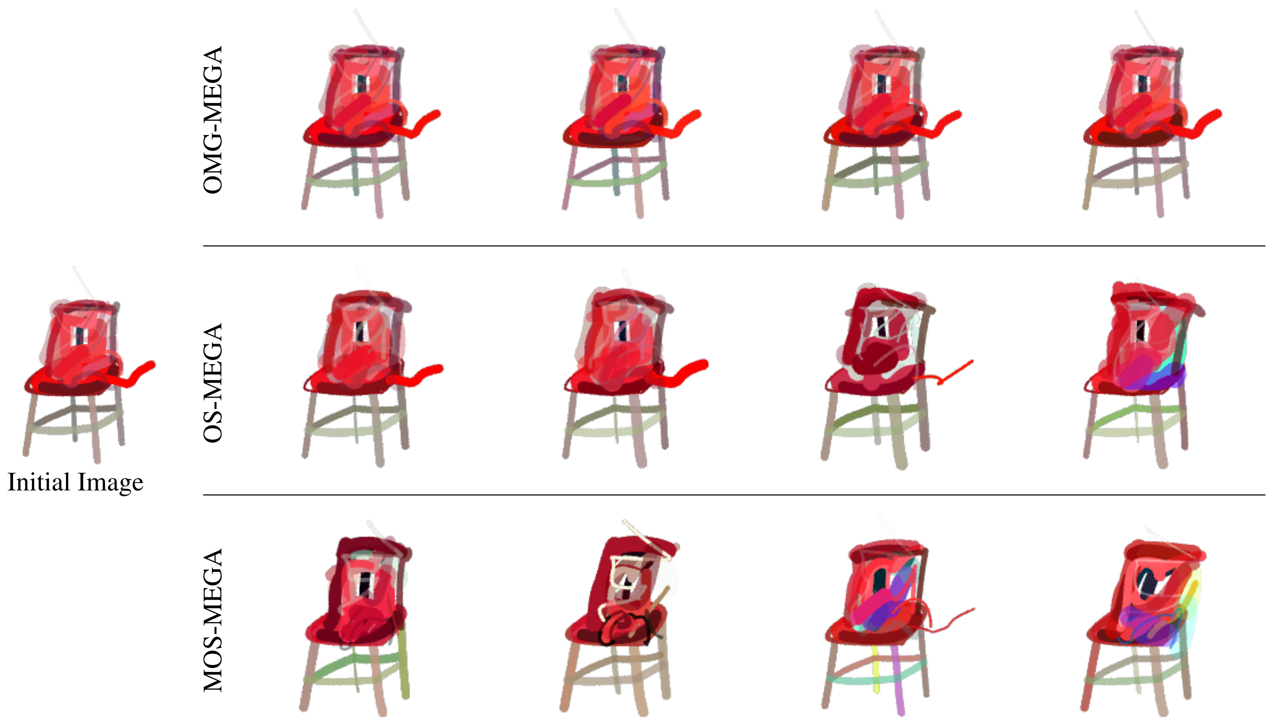


Figure 2: Each row shows four images randomly chosen from the population branched out from the initial image on the left, using one of the three algorithms.

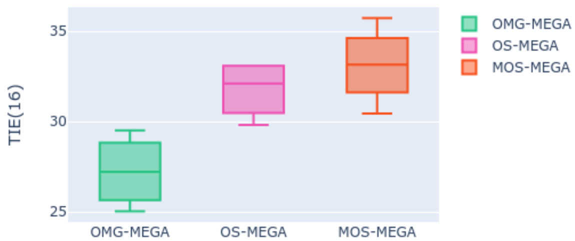


Figure 3: TIE values for the populations obtained with different algorithms. The experiments were carried using four different partial sketches.

Acknowledgments

We would like to acknowledge the Australian Research Council for funding this research (ID #DP200101059).

References

Dorst, K. 2015. *Frame innovation: Create new thinking by design*. MIT press.

Fontaine, M., and Nikolaidis, S. 2021. Differentiable quality diversity. *Advances in Neural Information Processing Systems* 34:10040–10052.

Gero, K. I.; Liu, V.; and Chilton, L. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*, 1002–1019.

Gero, J. S. 1998. Conceptual designing as a sequence of situated acts. *Artificial intelligence in structural engineering* 165–177.

Goldschmidt, G. 1991. The dialectics of sketching. *Creativity research journal* 4(2):123–143.

Grace, K.; Lawton, T.; and Ibarrola, F. 2023. When is a tool a tool? user perceptions of system agency in human-ai co-creative drawing. In *Designing Interactive Systems*.

Ibarrola, F.; Lawton, T.; and Grace, K. 2022. A collaborative, interactive and context-aware drawing agent for co-creative design. *arXiv preprint arXiv:2209.12588*.

Lawton, T.; Ibarrola, F. J.; Ventura, D.; and Grace, K. 2023. Drawing with reframer: Emergence and control in co-creative ai. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 264–277.

Li, T.-M.; Lukáč, M.; Gharbi, M.; and Ragan-Kelley, J. 2020. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics* 39(6):1–15.

Poon, J., and Maher, M. L. 1997. Co-evolution and emergence in design. *AI in Engineering* 11(3):319–327.

Pugh, J. K.; Soros, L. B.; and Stanley, K. O. 2016. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI* 40.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from

Figure 4: Randomly selected samples of sketches obtained with MOS-MEGA sampling for “a red chair”, “a hat”, “a lamp” and “a blue dress”, using “colorful” and “large” as behaviour dimensions.

natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Smith, G., and Whitehead, J. 2010. Analyzing the expressive range of a level generator. In *Proceedings of the 2010 workshop on procedural content generation in games*, 1–7.

Suwa, M.; Gero, J.; and Purcell, T. 2000. Unexpected discoveries and s-invention of design requirements: important vehicles for a design process. *Design studies* 21(6):539–567.

Williford, B.; Ray, S.; Koh, J. I.; Cherian, J.; Taele, P.; and Hammond, T. 2023. Exploring creativity support for concept art ideation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–7.

Can Creativity be Enhanced by Computational Tools?

Daniel Lopes*, **Jéssica Parente***, **Pedro Silva***, **Licínio Roque and Penousal Machado**
University of Coimbra, CISUC, DEI and LASI, University of Minho
{dfl, jparente, pedros, lir, machado}@dei.uc.pt
*These authors contributed equally

Abstract

The history of graphic design suggests that the introduction of new tools in people's workflow might promote new creative paths. This article discusses the impact of computational tools on performing creative tasks. We conducted semi-structured interviews with twelve professionals working on graphic design, data science, computer art, music and data visualisation. The results suggest scenarios in which it may be worth, or not, investing in developing creativity-enhancing tools.

Introduction

Looking backwards at the history of humanity or making a retrospective into our daily creative practices, it is noticeable that the use of different tools may introduce new creative possibilities. This study aimed to understand how creativity may be impacted by the introduction of computational tools, by studying and comparing non-computational, creativity support, e.g. desktop-publishing or photo-editing software, and Computational Creativity (CC) tools (co-creative and fully autonomous). To achieve this, we conducted semi-structured interviews with professionals working on Graphic Design (GD), data science, computer art, music and data visualisation. Most were experienced in more than one area. Despite including several creative areas, the research was primarily focused on GD. The answers were analysed and discussed to summarise the insights.

The remainder of this paper is organised as follows. The Background section presents a brief review of the literature regarding (i) creativity and the evolution of design tools (computational and non-computational); (ii) creative systems outside the design field; and (iii) studies on enhancing creativity. The Interviews section describes the research and discusses the results. Conclusions and Future Work summarise the work and present future research directions.

Background

Although the definition of creativity might not always be consensual, it is often accepted that novelty is one of the fundamental characteristics to define creativity (Boden 1996), and it may be achieved by exploring or extending the existing space of possibilities (Veale and Cardoso 2019). In the graphic design field, the introduction of the movable types by Gutenberg in the 1450s and the *Unigrid* system by Massimo Vignelli in 1977 (Philip B. Meggs 2016) are historical examples of extending the creative space by introducing new tools. The digital revolution recently brought new design tools and fostered new design movements (Lupton 2014). Additionally, software democratisation and easy-to-use coding libraries, e.g. *Processing (processing.org)*, fostered novel solutions such as animated and reactive designs (Shaughnessy 2012).

Concurrently, academics and practitioners started to explore Artificial Intelligence (AI) as a creative tool, establishing the Computational Creativity (CC) area — *"an emerging branch of* AI *that studies and exploits the potential of computers to be more than feature-rich tools, and to act as autonomous creators and co-creators in their own right."* (Veale and Cardoso 2019). CC tools may be co-creative or fully autonomous. The first ones collaborate with humans in creative tasks, while the second ones generate creative artefacts without human assistance (Maher et al. 2018). Nevertheless, both have been successful in aiding creativity, for instance, in areas such as computational art (Romero and Machado 2007; Machado et al. 2014; Elgammal et al. 2017), music (Miranda and Biles 2007; Farzaneh and Toroghi 2019; Loughran and O'Neill 2020) or design (Martins et al. 2016; Parente et al. 2020; Lopes, Correia, and Machado 2022), by applying evolutionary or machine learning techniques.

Besides art and design, creativity may be necessary for fields such as engineering. According to Robertson and Radcliffe (2009), engineers may be both positively and negatively influenced by creativity-support tools, since these may provide (i) a better ability to visualise and communicate ideas within the work team; yet these may also cause (ii) technical difficulties to make major changes in the projects as these get more complex; and (iii) limited creative possibilities imposed by the constraints of the tools. The authors argued that using computational tools may not be the best approach to generate ideas, yet these may be helpful to complement the human creative process.

Work on creativity-enhancing frameworks has also been done. Nickerson (1999) presented a framework composed of twelve steps for teaching creativity. For instance, (i) *"providing opportunities for choice and discovery"* or (ii) *"strategies for facilitating creative performance"*.

Shneiderman and Plaisant (2010) referred guidelines for developing creativity support tools, such as making them (i) *"low threshold, high ceiling, and wide walls"*, (ii) collaboration supportive, (iii) *"as simple as possible"* and (iv) able to *"iterate, iterate, then iterate again"*. CC tools may also fit in these guidelines, suggesting that their development may be desirable as well.

Furthermore, the interest in creativity-support computational techniques can also be noticed in the increasing number of creative coding classes in universities, schools and online courses (Dufva 2018; Hansen 2019).

Research Approach

This study aimed to understand (i) the impact of computational tools in creative tasks, mainly focusing on GD; and (ii) whether or not the insertion of new tools (computational or not) would enlarge the creative possibilities (opening new paths to explore in different directions). Assuming that new tools are favourable for enlarging the creative spectrum, computational tools might also do so. In that sense, creativity support and CC tools (co-creative and fully autonomous) were studied.

Semi-structured interviews guided by sub-questions were revealed to be an adequate method to address the goals of this study, providing clear strategies for organising data-gathering, coding and analysis. We conducted audio-recorded face-to-face interviews to address the nuances of the participants' language. Due to the nature of this research, only people who have worked with creativity-enhancing tools were included — 12 designers and computer artists (3 women and 9 men) working at the University of Coimbra (Portugal), from 26 to 61 years old with diverse backgrounds: (i) 3 seniors graphic designers; (ii) 2 senior CC researchers (iii) 3 PhD students researching on CC applied to graphic design; (iv) 2 PhD students researching on data visualisation; (v) 1 PhD student researching on data science and (vi) 1 PhD student researching on GD. The interviews took 15 to 30 minutes and were semi-structured by previously setting a list of 10 open-answer questions. If an answer responded to some further questions, we changed or skipped to avoid repetition.

Interview Analysis

To understand how computational tools may influence creativity, the research goal was decomposed into sub-questions: (i) do computational systems influence the creative process; (ii) is it worth investing in the development of creativity-enhancing computational tools; and (iii) how may CC tools be useful in the creative process. Therefore, these topics were organised under the following categories: (i) creative process and creativity; (ii) creativity-enhancing tools and their advantages; and (iii) CC tools.

Creative process and creativity

In the first questions of the interview, we aimed to understand the different backgrounds of the interviewees. Therefore, they were asked to describe the stages of their workflow and pinpoint the ones requiring creativity. It is important to highlight that the interviewees answered according to their own definitions of creativity.

From the content collected, the following common, fundamental stages were identified: (i) understanding the problem and the project requirements; (ii) searching existing work; and (iii) combining solutions for getting a new result. Additionally, it was consensual that the interpretation of the problem and prior experiences/knowledge (which may be influenced by the context one lives in) could affect the outcome. It was also consensual that all the stages of the workflow might require creativity.

Moreover, two interviewees argued that even searching may imply creativity, not only to find a better search method but also to find the best search domain. Two other interviewees believed that the most creativity-demanding stage is implementation, and one other claimed the requirements-gathering stage may be the one requiring less creativity. Moreover, it was assessed that creativity may also come from outside the work process. For example, by occasionally observing natural events or daily routines.

Creativity-enhancing tools and their advantages

The second group of questions was related to the use of computational and non-computational tools and aimed to assess: (i) which tools were used the most; (ii) whether and how these were helpful in the creative process; and (iii) how computational and non-computational tools may differ and in which contexts these may be used.

The answers revealed that all the interviewees frequently used computational tools in their creative process. Also, part of them claimed to use creativity support, version control and planning tools during the implementation phase. Most believed that such tools were highly advantageous, for example, by speeding up processes or fostering exploration, allowing otherwise unthinkable solutions. Also, some claimed that the introduction of computational tools brought control over the entire workflow, allowing one to go back and forward in the developments. One respondent argued that computational tools may provide a basis for starting or unlocking creative blocks, and others referred to the benefit of the internet in improving team collaboration and community support and providing easy access to new tools. Also, thanks to the easy access and the facility of creation, some declared themselves dependent on some tools.

Even so, most interviewees still use analogue methods, such as books for research or paper for fast sketching. Some noted that when using analogue methods, they need to better reflect on the execution process and exploration. Most of the interviewees added that the project and its needs may define the tools that are the most advantageous, and a PhD student working on CC claimed that the combination of computational and non-computational tools may be an asset to generate more experimental and less standard results.

Computational creativity tools

The final set of questions of the interviews regarding CC tools aimed to understand whether or not (i) these may be useful in the creative process, i.e. may one be inspired by a machine's outputs as one does by people's work; (ii) can people use these in real use scenarios; and (iii) is it worth the investment in research and development of such tools.

Most interviewees have expressed their interest in CC tools and believed that these may never replace human creativity, but complement it by increasing each others' capabilities. Nonetheless, there was a higher interest in co-creative tools over fully autonomous ones. Some admitted having used CC tools due to curiosity, to automate tasks or to access new functionalities, yet mostly to explore novel solutions.

From the above, one may infer that CC tools may foster new creative paths. Even so, some considerations were referred to: (i) such tools may be more effective on objective-evaluation issues; (i) CC systems may be picked or adapted according to the projects; (iii) most defended that humans will always guide the process. However, others claimed that having machines replace some human creative tasks may not be a negative thing, as people may adapt and direct their capabilities to more unexplored creative tasks.

Conclusions and Future Work

To collect perspectives on how computational tools may affect human creativity, we conducted semi-structured interviews with people working in creative fields such as graphic and computational design. The questions aimed to cover the creative background of the interviewees, understand which tools they use and for what purpose, and finally, collect their thoughts on CC tools. After coding and classifying the answers into themes, a further analysis was conducted for summing up the insights.

The answers revealed that the creative process may not be mainly shaped by the computational tools themselves but rather by social and personal background knowledge, which may change the interpretation of the problem.

However, especially in the early stages, the increasing productivity related to the use of new CC tools may be claimed as well-established evidence, as these may amplify the exploration and speed of the processes. Moreover, these may bring higher levels of confidence to the users by permitting them to revise and reformulate earlier developments without disabling further ones.

Also, the interviewees agreed that exploring new tools may expand creative possibilities, leading to new solutions. For instance, exploring both analogical and computational tools is recommended.

When questioning the role of CC tools in the creative process, the interviewees demonstrated their interest in co-creative tools and referred to their value for searching for unexpected solutions. Some divergence surfaced regarding fully-automatic tools due to the fear of human replacement. Others think it may be a natural way for humans to move their efforts forward to unexplored creative tasks.

In sum and paraphrasing one of the interviewees, all professions, processes of thinking and execution change and evolve in accordance with the evolution of their tools. Furthermore, personal background and experiences may have a strong impact on the employment of creativity, namely, due to social and cultural reasons.

Acknowledgments

This research was funded by national funds through the FCT — Foundation for Science and Technology, I.P., within the scope of the project CISUC — UID/CEC/00326/2020 and by the European Social Fund, through the Regional Operational Program Centro 2020 and is partially supported by *Fundação para a Ciência e Tecnologia*, under the grants: SFRH/BD/143553/2019, SFRH/BD/148706/2019 and SFRH/BD/144283/2019.

References

Boden, M. A. 1996. Creativity. In *Artificial intelligence*. Elsevier. 267–291.

Dufva, T. 2018. *Art education in the post-digital era - Experiential construction of knowledge through creative coding*. Doctoral thesis, Aalto University.

Elgammal, A. M.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the Eighth International Conference on Computational Creativity*, 96–103.

Farzaneh, M., and Toroghi, R. M. 2019. Music generation using an interactive evolutionary algorithm. In Djeddi, C.; Jamil, A.; and Siddiqi, I., eds., *MedPRAI*, volume 1144 of *Communications in Computer and Information Science*, 207–217. Springer.

Hansen, S. 2019. public class graphic design implements code {Yes, but how?}: an investigation towards bespoke creative coding programming courses in graphic design education.

Lopes, D.; Correia, J.; and Machado, P. 2022. Evodesigner: Evolving poster layouts. *Entropy* 24(12).

Loughran, R., and O'Neill, M. 2020. Evolutionary music: applying evolutionary computation to the art of creating music. *Genetic Programming and Evolvable Machines* 21(1-2):55–85.

Lupton, E. 2014. *Beautiful Users: Designing with User-Generated Content*. Princeton Architectural Press.

Machado, P.; Martins, T.; Amaro, H.; and Abreu, P. H. 2014. An interface for fitness function design. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design – Third International Conference, EvoMUSART 2014. Proceedings*, volume 8601 of *Lecture Notes in Computer Science*.

Maher, M. L.; Grace, K.; Karimi, P.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. In *Proceedings of the Ninth International Conference on Computational Creativity*, 104–111.

Martins, T.; Correia, J.; Costa, E.; and Machado, P. 2016. Evotype: from shapes to glyphs. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 261–268. USA: ACM.

Miranda, E. R., and Biles, J. A. 2007. *Evolutionary Computer Music*. Berlin, Heidelberg: Springer-Verlag.

Nickerson, R. S. 1999. Enhancing creativity. *Handbook of creativity* 392.

Parente, J.; Martins, T.; Bicker, J.; and Bicker, J. 2020. Which type is your type? In *Eleventh International Conference on Computational Creativity*.

Philip B. Meggs, A. W. P. 2016. *Meggs' history of graphic design*. John Wiley & Sons.

Robertson, B., and Radcliffe, D. 2009. Impact of cad tools on creative problem solving in engineering design. *Computer-aided design* 41(3):136–146.

Romero, J., and Machado, P., eds. 2007. *The Art of Artificial Evolution: A Handbook on Evolutionary Art and Music*. Springer.

Shaughnessy, A. 2012. When less really does mean less: Design observer.

Shneiderman, B., and Plaisant, C. 2010. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

Veale, T., and Cardoso, F. A. 2019. *Computational creativity: The philosophy and engineering of autonomously creative systems*. Springer.

Exploring post-phenomenological perspectives on creativity with GPT-4
Paper type: CC bridges

Giovanni Lion
School of Design
The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom, Hong Kong
giovanni.lion@gmail.com

Abstract

This paper explores the use of post-phenomenology, a contemporary philosophical movement, as a framework for understanding the relationship between humans, technology, and creativity. The theory of mediation proposed by Don Ihde is employed to categorize different types of technological relations that shape creative contexts. Large Language Models (LLMs) are used to generate post-phenomenological descriptions of arbitrary examples of technology. The generated formalizations can be evaluated and refined theoretically, contributing to the evaluation of the theory's applicability to various creative scenarios. This *remedial* attitude in proceeding emerges as a methodological bridge between computational creativity and design. Furthermore, reflecting on this process of inquiry highlights the inherent creative potential of LLMs and their application in exploratory practices.

Introduction

Post-phenomenology, an interdisciplinary and contemporary philosophical movement, has emerged as a significant framework in design research for understanding the complex relationship between humans and technology (Ihde 1990; Verbeek 2015; Latour 2008; Benjamin et al. 2021). The Computational Creativity (CC) community has been critically examining the role technology has in shaping our creative processes and products, and one potential avenue to approach this examination is through the lens of post-phenomenology. By adopting a post-phenomenological interpretation of technology, the CC community can gain a deeper understanding of the ways in which humans and non-human entities interact creatively with technology. This interpretation recognizes that our interactive experiences with the non-human are not solely determined by the technology, but also by the contexts, practices, and social norms in which they are embedded. Ultimately, developing a post-phenomenological view of technology can help the CC community better understand and engage with the complex and dynamic relationships between humans, technology, and creativity.

Don Ihde's work provides a useful framework for understanding the complex relationships between humans, technology, and the world. According to Ihde (1990), the idea of *mediation* could be schematized as follows:

- Unmediated perception: I—World
- Mediated perception: I—Technology—World

He identifies four fundamental types of relations that characterize these interactions, summarized in Table 1. The first type is *embodiment* relations, in which technologies unite with a person and point their unity outward at the outside world. Examples of this type of relation include using a phone to talk to other people or viewing objects through a microscope. In *hermeneutic* relations, people interpret how technologies reflect the world, such as reading an MRI scan or using a metal detector to detect metal. In *alterity* relations, people engage in technological contact with the outside world acting as a backdrop, as in the case of interacting with robots or using ATMs. Finally, Ihde distinguishes between *background* connections and technologies that frame human experiences and behaviors, such as the sounds of air conditioners or notification sounds from cellphones during

Name	Form	Definition
Mediated Embodied	(I — T) → W	Broaden the area of sensitivity of our bodies to the world (e.g. glasses, a dental probe, a paintbrush)
Mediated Hermeneutic	I → (T — W)	Provide a representation of the world that we need to interpret (e.g. thermometer, watch)
Alterity	I → T (— W)	Humans are related to or with technology as a quasi-other (e.g. ATMs, robots)
Background	I (— T / W)	Shapes the context of our experience in a way that is not consciously experienced (e.g. refrigerators, central heating system)

Table 1: A summary of the relations types proposed by Ihde and their formalization. In the examples in the second column, I represents the human, T stands for Technology, and W refers to World.

a conversation. By understanding how these different types of relations shape the creative context, it is possible to gain better insight into how technology can be used to support and enhance the creative process.

The theory also comes with notation system defined as follows:

- **—** simple connections between entities
- **→** interpretation of one by the other
- **( )** being experienced together
- **/** being in the background of another entity
- **[ ]** being already contextualized before being processed

This paper explores the explanatory capabilities of mediation theory by employing GPT-4 to generate formalizations of arbitrary instances of technology. Starting from a prompt containing the theory and some examples, GPT-4 is asked to analyze the use of a technology and return the post-phenomenological notation for the mediation. On one hand, this approach contributes to evaluate the theory's applicability to various creative scenarios. On the other, it showcases the potential of Large Language Models (LLMs) as a creative tool in academic research and analysis.

Towards a post-phenomenology of computational creativity

Embodied mediation is perhaps the stereotypical form of creative technology use. We imagine a painter with a brush in their hand, a musician with their instrument and a writer with their pen. However, what seems to characterize the mediation in creative endeavors is the non-utilitarian context of the action, holding the stage in the *background* (/). Consider the example of a pencil. It can be used to write a grocery list or to draw a sketch. In post-phenomenological terms we could describe the two forms of mediation as:

1. $(I - Pencil) \to GroceryList$
2. $(I - Pencil) \to Sketch$

However, these formalizations fail to capture the different contexts in which the mediation takes place. Contextualizations, represented by $C_n[]$, can help us identify a specific framing that is applied to the mediation's conceptual input. For example we could expand the examples as:

3. $(I - Pencil)/C_1[Items] \to GroceryList$
4. $(I - Pencil)/C_2[Subject] \to Sketch$

where C_1 is most likely a *shopping* context, while C_2 is the domain of *visual artistic expression*. To better understand these formalizations, it may be useful to imagine how the same contextualizations might be applied to a different output:

5. $(I - Pencil)/C_1[Items] \to Sketch$
6. $(I - Pencil)/C_2[Subject] \to GroceryList$

Mediation in sentence 5 could represent a sketch of items used as shopping list, while sentence 6 would suggest an artistic piece, perhaps a poem, produced as a list of groceries.

In CC literature there are numerous examples of systems affording mediation spanning all forms of relations, but the discussion about non-human creativity roughly coincides with discussing the notion of *alterity* in post-phenomenology. Frieder Nake, a pioneer in *generative art*, already saw this as a critical aspect of his type of work when he said: "the individual human subject simply did not exist anymore, once he or she had set the boundary conditions for the image to be computed". Nake repurposed plotters, originally used for technical drawing and used them to generate artistic designs based on generative algorithms. We can frame Nake's new formed relationship with the plotter (8) in comparison with its generic form (7)

7. $I \to Plotter(-TechnicalDrawing)$
8. $I \to Plotter/Art[Nake'sAlgorithm](-Designs)$

Explorative prompting

A technological relation is not immediately obvious to the observer and in some cases its formalization may be ambiguous and in need for clarification. In order to explore how this theory might represent different instances of technological mediation it is possible to test its explanatory power by evaluating specific examples.

LLMs, such as GPT-3.5 and GPT-4 have shown the ability to manipulate relatively formal language compositionality, such as the one adopted by Ihde's theory. In an attempt to produce plausible formalizations, I explore the use of LLMs to generate post-phenomenological descriptions, starting from a prompt containing a brief summary of the theory and some examples.

Prompt	**GPT-4 response**
Car	I → Car (— Transportation)
Computer	I → Computer (— Data Processing)
Caliper	I → (Caliper — Measurement)
Railway system	I (— Transportation / Railway System)
Joystick	(I — Joystick) → Game Control
Electrical plug	I (— Electrical Plug / Power Supply)
Door	I → Door (— Access Control)
Door handle	(I — Door Handle) → Door Opening
Operating a door handle which is part of an art installation	(I — Door Handle) / Art[Installation] → Interaction

Table 2: This table contains GPT-4 responses to the user message specified in the first column.

The examples in Table 2 show that the prompt used is sufficient to allow simple technologies to be accurately described. From this starting point, it is then possible to challenge GPT-4 with more complex technologies. For example, when prompted with the abstract of Colton's paper describing "The Painting Fool" (Colton 2012b), it simply returns

9. $I \rightarrow ThePaintingFool(-CreativeArt)$

It is possible to ask more explicitly to extend a particular mediation, in order to explore possible expansions. Table 3 shows how to prompt the model to follow a particular form when generating the expansion. However, GPT-4 does not perform this kind of instructions reliably: the generated items put the contextualization before the interpretation symbol (→), while the prompt explicitly specifies the opposite order.

User: I → The Painting Fool / *Background → *Contextualization[*Input] (— Artistic Creation) Replace the elements starting with * in this formula, introducing background relations, contextualizations and their inputs. Give me three examples.
I → The Painting Fool / Art History[Paintings] → Interpretation (— Artistic Creation)
I → The Painting Fool / Color Theory[Palette] → Selection (— Artistic Creation)
I → The Painting Fool / Human Preferences[Feedback] → Adaptation (— Artistic Creation)

Table 3: Background expansions on "The Painting Fool".

The bot can also be asked to explain verbally its own examples of background relations expansions, as shown in the conversation presented in Table 4. These generated relations can then be evaluated and potentially further explored and refined theoretically.

It is important to note that in this process there are no expectations of GPT-4 producing *correct* answers. In fact, inaccurate formalizations are perhaps more valuable in a theory building context because they can highlight ambiguous elements that require deeper inquiry. Attempting to evaluate GPT-4's output is in itself an act of reflection (Schön 1991) over the theory and its operationalization. For example, by looking at the background relations described in Table 4 some considerations may arise. First, the data-driven nature of LLMs is not represented explicitly. Second, all the results presented by GPT-4 present the background relation attached to the I, before the interpretation step (→), which is unexpected, but also meaningful. The ordering is different from the one used in Table 3; this form might emphasize the way our existing background relations affect our expectations about a given technology. These considerations are only possible because GPT-4 *deviated* from the instructions and examples in the system prompt.

Being presented with incomplete or incorrect examples is conducive to deeper and more precise reflections that can lead to theory improvements. It may be possible to generalize this phenomenon to a case of mediation where generative tools are used in a context where *mistakes* are not penalized. In such situations, divergence might be desirable because errors and misclassifications constitute a starting point to improve the process.

User: GPT-4 used for creative writing
I → GPT-4 (— Creative Writing)
User: Can you expand the formula highlighting some of the major background relations and contextualizations involved? Give me three examples of expansions of this formula I → GPT-4 (— Creative Writing)
Sure! Here are three examples of how the formula I → GPT-4 (— Creative Writing) can be expanded to include background relations and contextualizations: 1. I (— Writing / Literature) → GPT-4 (— Creative Writing) In this expansion, the background relation is the long history of writing and literature that has shaped human creativity and expression. GPT-4 is contextualized within this history as a new tool for creative writing, building on the legacy of human writers and their works. 2. I (— Technology / AI) → GPT-4 (— Creative Writing) Here, the background relation is the broader context of technology and artificial intelligence that has enabled the development of GPT-4. The formula highlights the role of GPT-4 as a product of human ingenuity and technological progress, and its potential to transform the field of creative writing. 3. I (— Culture / Society) → GPT-4 (— Creative Writing) In this expansion, the background relation is the cultural and social context in which creative writing takes place. GPT-4 is situated within this context as a tool that reflects and shapes cultural values, norms, and practices. The formula highlights the potential of GPT-4 to challenge existing cultural assumptions and to open up new possibilities for creative expression.

Table 4: Background relations expansions on "GPT-4 used for creative writing".

As proposed by Hoorn (2023), probabilistic models can be considered inherently creative because they are error-prone. Expanding on Hoorn's account of text-to-image models, it seems that LLMs also allow for a chance to encounter unexpected variations and broken results that can inform further interaction and adjustments. This iterative process is very familiar to design practices, as Bruno Latour suggests "[Design] is never a process that begins from scratch: to design is always to redesign. There is always something that exists first as a given, as an issue, as a problem" (2008). Boden's account of *exploratory* creativity (1992) might perhaps fit Latour's description of design. The argument in favor of data-driven generative tools in this context is that they do not always work as expected, therefore allowing explorations beyond the boundaries of mental fixation.

A bridge from CC to design

The relationship between CC and design is a crucial aspect to consider when examining the broader implications of post-phenomenology in creative practices. Design, as an inherently creative field, involves problem-solving, decision-making, and generating novel solutions. Design researchers are often interested in evaluating practical applications of CC systems (Kim 2023; Liu 2022) or investigating the interaction with as *alterity* relation (Algarni 2020; Ragot 2020). Yet, I believe there is value in creating a better theoretical understanding of how technological mediation takes place in a creative context.

For example, understanding how *background* relations shape the expectations attached to a creative system warrants a deeper comprehension of the technology used, its inner components and the social environment in which is embedded. Forming this multi-faceted perspective requires bridging knowledge about output and process evaluation, topics that are thoroughly discussed in the CC community (Jordanous 2012; Wiggins 2006; Colton 2012a), with considerations about technological interaction and its impact on society, which is a topic central to design.

According to Latour (1990), technology is what makes society durable, as a purely social world cannot exist. Stability is generated through the assemblage of a diverse network of humans and non-humans. Latour illustrates this by explaining that a door is a prime example of a heterogeneous network. If a door were to be removed, a lot of effort would be required by the human to achieve the same purpose. To enter, a new hole would need to be made and then bricked up. However, with the door, both the human and non-human can work together to allow entry. The door must be presented in a way that it can be opened, and the human must interact with it in a specific way to open it. Latour believes that the symmetry of this interaction is what creates stability in society.

This analogy applied to a creative context is extremely powerful, as it allows for a non-dual way to look at creative practices mediated by technology. If the mediation happens in a creative context, the contextualization of the technology becomes crucial in shaping the expectations and outcomes of the creative process. This means that the technological tools used in a creative project are not neutral but actively shape the creative process and its outcomes. Therefore, understanding the technological mediation in a creative context requires a multi-disciplinary approach that encompasses both CC and design perspectives.

The intersection of CC and design provides a valuable opportunity to study the role of technological mediation in creative practices. Understanding how technology shapes creative processes and outcomes requires a multi-disciplinary approach that encompasses both output and process evaluation from the CC perspective and considerations about technological interaction and its impact on society from the design perspective. By bridging these two fields, a better theoretical understanding of technological mediation in a creative context can be achieved. This understanding is critical in developing more nuanced and effective approaches to leveraging technology for creative endeavors.

Limitations

While the post-phenomenological framework offers valuable insights into understanding the complex relationships between humans, technology, and creativity, it also has its limitations. One limitation is that the framework is primarily descriptive and does not provide clear guidelines on how to design or evaluate creative technologies. Moreover, the formalization of human-technology relations using Ihde's notation system may oversimplify the intricate dynamics that occur in real-world creative scenarios, and may not fully capture the nuanced and context-dependent nature of these interactions. Further investigation is needed to form a more systematic and rigorous methodology to evaluate the post-phenomenological framework in the context of computational creativity and design.

Another limitation lies in the epistemological status of post-phenomenological claims. Without a coherent methodology, post-phenomenology may not provide replicable and consistent findings that possess predictive power, essential for scientific advancement. Drawing from post-modern thought and potentially introducing relativism into research outcomes, post-phenomenological investigations may struggle to deliver lasting and transferable insights. To ensure more robust scientific exploration, a better understanding of the role of language in human-technology interactions and the integration of more empirical language approaches may be required (Smith 2014).

Conclusions

This paper explored the potential of post-phenomenology as a framework for understanding the relationship between humans, technology, and creativity. By being exposed to the post-phenomenological interpretation of technology, the CC community might develop a deeper understanding of the different ways in which the human and non-human can interact in creative practices. The theory of mediation proposed by Don Ihde was used to produce descriptions of interactions using GPT-4. The generated formalizations and their shortcomings showed potential directions for improvement of the theory. Furthermore, the exploration of the theory's explanatory power using GPT-4 highlighted the importance of *background* relations and contextualizations in shaping technological mediation in creative practices. The relationship between CC and design was also discussed in the context of post-phenomenology, showing synergies and similarities between disciplines, which may foster opportunities for collaborations.

References

Algarni, A. 2020. Neuroscience of creativity in human computer interaction. In Arai, K.; Bhatia, R.; and Kapoor, S., eds., *Proceedings of the Future Technologies Conference (FTC) 2019*, 248–262. Cham: Springer International Publishing.

Benjamin, J. J.; Berger, A.; Merrill, N.; and Pierce, J. 2021. Machine learning uncertainty as a design material: A post-phenomenological inquiry. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI

'21. New York, NY, USA: Association for Computing Machinery.

Boden, M. 1992. *The creative mind*. London: Abacus.

Colton, S. 2012a. Computational creativity: The final frontier? volume 242, 21–26. IOS PressNieuwe Hemweg 6BAmsterdam1013 BG.

Colton, S. 2012b. *The painting fool: Stories from building an automated painter*, volume 9783642317279. Springer-Verlag Berlin Heidelberg. 3–38.

Hoorn, J. F. 2023. Computer-vision classification-algorithms are inherently creative when error-prone. In *Proceedings of the 18th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry*, VRCAI '22. New York, NY, USA: Association for Computing Machinery.

Ihde, D. 1990. *Technology and the lifeworld*. Indiana University Press.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4:246–279.

Kim, J. 2023. The effect of AI-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation* –.

Latour, B. 1990. Technology is society made durable. *The Sociological Review* 38(1_suppl):103–131.

Latour, B. 2008. A cautious prometheus?: A few steps toward a philosophy of design (with special attention to peter sloterdijk). *Annual International Conference of the Design History Society*.

Liu, V. 2022. Design guidelines for prompt engineering text-to-image generative models. –. Association for Computing Machinery.

Ragot, M. 2020. AI-generated vs. human artworks. a perception bias towards artificial intelligence? In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery.

Schön, D. A. 1991. *The reflective practitioner*. Taylor & Francis Ltd.

Smith, D. 2014. Rewriting the constitution: A critique of 'postphenomenology'. *Philosophy & Technology* 28(4):533–551.

Verbeek, P.-P. 2015. Beyond interaction: A short introduction to mediation theory. *Interactions* 22(3):26–31.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

What Makes Gameplay Creative?

Brad Spendlove[1] and Dan Brown[2]
[1] Computer Science Department, Brigham Young University, USA
[2] Cheriton School of Computer Science, University of Waterloo, Canada
brad.spendlove@byu.edu, dan.brown@uwaterloo.ca

Abstract

We consider the question of where the threshold is for the play of a game to be creative. The question of creativity is connected to, but not the same as, the complexity of finding optimal play; in particular, just because a game is *hard* to play well, that does not inherently make it a *creative* game to play. To clarify that difference, we introduce a set of desiderata for determining whether playing a given game is a creative task. Inspired by a recent paper, we examine the word game Codenames as an example of how these desiderata can be applied to analyze whether a game admits creativity. Our overall goal is to explore the relationship between creativity, algorithmic gameplay and fun in games that are clearly ones in which a computer could be a player.

Introduction

What properties make the play of a game a creative task? Some games seem clearly creative to play: for example, playing Charades or Pictionary requires pantomime or drawing; other games involve competitive storytelling or coming up with songs or other creative artifacts. Other games likely are not creative: games of pure chance certainly are not, but neither are games in which the optimal choice of action for a player is clearly defined by a simple algorithm, such as Blackjack or Connect Four. In general, the process a player must engage with in developing either turn-by-turn tactics, or the overall strategy that they are implementing, must have properties consistent with creativity for gameplay to be a creative task.

Humans are not the only game players, of course; pets play games with their human companions, and non-human primates play games together (Kaufman and Kaufman 2015). When animals play games, they also sometimes discover novel, successful strategies, and in doing so, they also can be creative participants in a game. Similarly, agents like computers can also be game players, and the strategies and approaches they discover in their gameplay can also be novel and successful, and hence creative. Gameplay is thus one entertaining route into the overall field of creativity studies in general, and computational creativity in particular.

Our key idea is that creativity is a facet of a rule set: a game either does or does not admit creative play by its nature, and modifying a game's rules may change the game's creativity status. We identify a set of desiderata that we believe are necessary for games to be creative.

Gameplay and creativity

It is easy to categorize games at the extreme ends of the spectrum of possible creativity they allow. Famous chess and Go players are often characterized by gameplay that is universally praised as creative, and it is difficult to imagine arguments in favor of tic-tac-toe allowing for creativity. But is there space for creativity in, for example, checkers? This potential uncertainty leads to two questions. First, what qualities of a given game lead us to categorize it as allowing creativity or not? And second, how can those qualities be applied to games whose status as creative is less clear? Answering these questions can also help to clarify relative differences between creativity in games. Checkers is strategically deeper than tic-tac-toe, but is it deep enough to allow for creativity? A robust framework for categorizing creativity should be able to account for the difference between strategy and creativity.

Natural language games

It is instructive to consider the case of natural language games. The richness of language is and has been a constant source of joy in human life. Wordplay, rhyming, jokes, songs, and poetry all draw on the depth, emotion, and shared understanding afforded by language. Language is powerful; it must be to describe the complexity of our reality, both internal and external. And within that power and complexity is an inexhaustible combinatorial space of words, forms, and ideas. The clever navigation of that space to find novel and meaningful expressions is a prime example of creativity.

The rules of a word game, like the forms of poetry or song lyrics, establish a shared portion of the language space for the players to explore together. And while artistic expressions may be difficult to evaluate, a game's rules provide a common metric to compare and judge plays. These traits are desirable from a computational creativity perspective. Word games can be creative domains with a constrained combinatorial space and clear evaluation criteria, which serve as useful footholds for designing, improving, and sharing creative computer systems that play them.

Such games clearly allow for (or even require) creativity to play, and we can examine their properties in terms that can be applied to non-language games to usefully reason about the creativity they permit.

Characterizing Creative Games

We propose that the degree to which a game admits creative play can be characterized by examining the space of possible moves in the game, how that space changes as the game develops, and the algorithm that decides the winner of the game. Specifically, we say that a game admits creative play—i.e. playing the game is a creative act—if it has a sufficiently large space of possible moves that changes meaningfully as the game progresses. We also consider whether the game's outcomes can be decided by a highly compressible algorithm. This last criterion does not affect the creativity admitted by a game, but a game that satisfies it falls into a category of creative games that is particularly interesting to computational creativity.

Large set of possible moves

We first draw a distinction between the game-theoretic complexity of a game and the space of possible moves the player may make in the game, the latter of which pertains to the current discussion. Game complexity can be reasoned about by metrics pertaining to search strategies and computational complexity. One such measure is the total number of possible games for that set of rules. Tic-tac-toe is an example of a game with a very small number of possible games that are trivial to exhaustively explore by a computer or adult human. It is not widely considered to admit creativity. Conversely, chess is a game with approximately 10^{40} sensible games[1] and has been demonstrated to admit creativity.

Reasoning about the number of possible games is not sufficient to determine whether a game admits creativity, however. We can construct pathological examples of nonsense games that technically have a large number of possible games. For example, consider a "googol guessing game" where a random number between 1 and 10^{100} is generated and the player has one chance to guess it. There are a very large number of such games, but they are not interesting and playing them is not creative. Therefore, in addition to the size of the game space, we also consider the (potentially) creative task that faces the player at each ply.

Our first desideratum is that the space of possible moves at any given ply is large enough to admit creativity. When framing a game ply as a creative task, we can reason about how many possible actions the player could take. We may count the full space, restrict it to sensible moves, or include hidden information in the formulation of the creative artifact.

Meaningful difference between possible moves

The combinatorial size of a game alone is not sufficient for a game to admit creativity—recall the googol guessing game. Thus, our second desideratum is that the game rules give rise to meaningfully different sets of choices at each ply. We say that difference is meaningful if either the best strategy for selecting a move, the goal of the current move, or the space of moves itself differ from ply to ply. In other words, the space must accommodate artifacts that are recognizably different from one another and are of varying levels of quality. We discuss artifact novelty and quality further in a later section.

This desideratum rules out the googol guessing game as one that admits creativity because the strategy is only ever to guess a random number. Playing that game is not creative because each outcome is functionally identical. Modifying the game to allow for an unlimited number of guesses is similarly uninteresting. If we further modify the game to say whether a guess was higher or lower than the target number, the strategy to search the space is an obvious binary search. To admit creativity, the task of playing the game must present a meaningful creative task at every ply.

Chess, again, is an example of a game that admits creativity. There are many possible games of chess, but crucially the way the game plays out is almost guaranteed to result in a game that has never been played before. Thus, at each ply, the player is likely to face a completely novel set of options, each with the potential to produce further novel game states. In this way, the game rules give rise to effectively limitless creative tasks with different spaces of possible moves. This is also an example of a common way to inject variety into a game: by pitting opponents against one another. When two players are changing the game state to achieve opposing goals, the game can give rise to many different problems for players to solve.

Deterministically decidable games

What separates games from other creative tasks is that games have concrete win and lose states. One player wins, and the other(s) lose. Even if a game is not zero-sum, there is still a rule-defined goal and some method for determining which player has a higher score or equivalent measure.

As we are computational creativity researchers, one key goal for us is to describe creative games that computers can themselves play, and where the winner of the game is easily determined. As such, we seek rule sets for which there exists a straightforward description of the current state of a game and an easy-to-describe algorithm that, given the state of the game, can identify the winner.

We can divide games into two categories: games whose outcomes are determined by a deterministic, low-complexity algorithm and those that are not[2]. Examples of the former are myriad and include chess and checkers as well as complex war games. Matches of even the largest of such games can be easily represented as a list of moves taken, and the algorithm to decide the outcome of the game consisting of those moves is trivial.

Even real-time video games ranging from Super Mario Bros. to modern first-person shooters are deterministic over a set of inputs and starting game state and have relatively simple game state representations (that may be embellished

[1]This is a modification of the well-known Shannon number—10^{120}—which assumes there are 30 legal moves at each ply over a game of 80 plies (Shannon 1950). Grime (2015) instead substitutes 3 as the approximate number of "sensible" moves at each ply.

[2]The authors of this paper are not aware of any popular games with a deterministic outcome but a highly computationally complex decision algorithm.

by entertaining graphical representations). This is evidenced by the efficiency of multiplayer network code and the existence of shareable sequences of inputs for recreating gameplay on another computer, e.g. tool-assisted speedruns or Doom demos (Lowood 2008).

The other class of games is those that are not deterministically decidable by an algorithm, either because of a game state that is not tractably representable or a subjective scoring system. Examples of such games include artistic competitions, sports, and some board games such as Apples to Apples (Kirby and Osterhaus 2007). Due to their complexity, these games often fulfill the desiderata for games that admit creativity. However, we propose that algorithmically decidable games that admit creativity are especially interesting to computational creativity research.

Such games provide unique opportunities to study creative tasks that have outcomes of deterministically decidable quality, i.e. their contribution toward winning or losing the game. Evaluating the quality of a creative artifact is a central challenge in computational creativity and is often very difficult. Free-form creative domains in the arts often have no agreed-upon subjective evaluation criteria among human critics, let alone computationally tractable ones.

If a game satisfies the two desiderata presented herein, we argue that playing the game qualifies as a creative task. The creative agent fulfilling the task is working within a large space of possible artifacts that compare meaningfully with each other, the agent's creative responsibilities inform how the artifact space is searched, and both the agent and their audience can evaluate the quality of an artifact.

Connection to Creativity Theory

In this section, we explore how the desiderata for creativity-admitting games that we have presented relate to existing theories of creativity. We will demonstrate that they reflect important considerations for reasoning about creativity.

Ritchie (2007) introduces three necessary qualities of an artifact that is to be considered creative: novelty, quality, and typicality. If an artifact exhibits these properties, then it follows that the agent responsible for the artifact's creation behaved creatively. As we are evaluating game rulesets instead of single artifacts, we instead interrogate whether the game's task can produce artifacts with these qualities. In other words, we say that the space of possible moves at a given ply admits creativity if the artifacts in that space can be meaningfully novel, high-quality, and typical.

The relationship that deterministically decidable creative games have to these qualities makes such games notable and interesting for computational creativity research. Under this theory, creativity requires novel, high-quality, and typical artifacts. Games enforce typicality through the rules and social contract the players enter into when playing the game. We have discussed how deterministically decidable games represent a uniquely tractable means of evaluating quality. Our desiderata of a large space of meaningfully different possible moves reflect whether the task of playing the game allows for novelty and quality in its output artifacts.

For a counterexample, consider the googol guessing game. Of all the guess artifacts that comprise the space, one is correct and the others are incorrect. Artifacts in the space do have different quality measures, but the space admits neither the richness nor nuance of differences between the quality of two given artifacts that characterize creative domains. Worse still, all artifacts in the space have the same novelty: they are all equivalently uninteresting guesses. No matter how many artifacts the game player generates, none of them are novel, and therefore the act is not creative (Colton, Pease, and Ritchie 2001).

Instead of focusing on qualities of creative artifacts, the creativity tripod described in Colton (2012) describes three capabilities that an agent must necessarily display to be judged as creative: skill, appreciation, and imagination. Through this lens, we can reason about whether a given game task requires these capabilities and allows an agent to express them. Appreciation is the agent's skill that allows them to evaluate the quality of an artifact. Thus, just as a quality measure is implicit in the rules of a deterministically decidable game, that same decision algorithm can be run by the agents playing the game. Thus the game requires and exercises the agent's appreciation.

Skill is a requirement for playing all but the most simple of games, whether they admit creativity or not. The more difficult or complex the game, the more skill is required. Many games that do not admit creativity still require skill, such as agility- or precision-based competitions. Even executing a known strategy in a solved game such as checkers (Schaeffer et al. 2007) can be considered skillful if that strategy is complex enough. Successfully navigating a space of possible actions large enough to admit creativity is certainly a skillful endeavor. Highly strategic games that are not open or complex enough to admit creativity—which are not the focus of this work—fall somewhere in the middle of a spectrum of games that require low skill, to games that require high skill, to games that admit creativity.

Along similar lines, we may consider imagination as a factor that distinguishes between strategic play and creative play. Although it may take skill to execute a complex strategy, it by definition does not require imagination. Inventing new approaches to problems or finding especially clever lines of play are only possible in a game that satisfies our desiderata and admits creativity. An agent must have imagination to successfully play such a game.

Example: Codenames

We can apply our desiderata to games to examine the degree to which they admit creativity. We will use Codenames (Chvátil 2015) as an example of this analysis.

Spendlove and Ventura (2022) presented Codenames as an example of a creative language game and claimed that playing the spymaster role in the game was a creative task. Codenames is a game of communicating secret information via one-word clues. The game is played with two teams using a grid of 25 word cards dealt from a large deck. One member of each team is the spymaster, who can see a secret key that shows which of the word cards belong to their team, the other team, or neither team. Teams take turns with the spymaster giving a one-word clue related to the team's word cards and a number that signifies how many such cards the spymaster intends the clue to relate to. The rest of their team then guesses one word card at a time, and its secret role is revealed. Any incorrect guess ending the team's turn. The goal is for each team to identify all of their cards before the opponents identify theirs.

The spymaster's task is to come up with a clue word that relates to some subset of their team's word cards while *not* relating to any of the other cards. This can be represented as a graphlet with connections between a potential clue word and any word cards it relates to, positively or negatively. Because the clue word can be any English word, there is a very large number of such graphlets the spymaster must consider. Furthermore, as the game progresses and some word cards are guessed, they are removed from the set under the spymaster's consideration, changing the creative task from ply to play. Codenames, therefore, fulfills our desiderata for admitting creativity. Many different sets of targeted word cards and clues can be selected at any given ply, and clues can be obvious (likely relating to only one word card) or surprising. High-quality clues will lead the spymaster's teammates to correctly guess the intended word cards, while low-quality clues will result in fewer or no correct guesses.

Additionally, Codenames is decidable by a very simple algorithm. Regardless of the complexity of the task of selecting a clue, a team's turn consists of a clue and a number of guesses that update the game state, both of which are simple to represent. Given a starting state and a history of guesses, it is trivial to determine both who wins the game and the useful intermediate measure of how many of each team's cards remain. Thus, Codenames also fulfills the additional decidability criteria. By this analysis, Codenames is indeed a creative task with a tractable evaluation metric that merits further computational creativity research.

Future Work

Our desiderata for determining creative gameplay can serve as tools to interrogate a spectrum of games and rulesets. Most notably, we see potential in examining how the creativity admitted by a game changes as its rules are changed. Identifying a cross-over point in a series of game tasks that share similar rules could provide more specific insights into how rules shape play spaces.

Similarly, there is potential for more granular analysis of games as they naturally evolve over the course of play. For example, chess openings and endgames can be memorized and solved to some extent, but the middlegame still represents a space of possible moves large and complex enough to admit creativity.

Finally, this work could serve as a useful tool for game designers. A legitimate goal for a game designer is to design a game that admits creative play. Having a framework for reasoning about the creativity admitted by a game could aid game designers in analyzing and improving such games.

Conclusion

Games may represent a fruitful vein for computational creativity research; their concrete win and lose states could serve as a unique foothold for evaluating creative artifacts. However, because games differ from more traditional creative domains, it may be unclear whether playing a given game is truly a creative task.

In this paper, we have described considerations for how creativity in games can be analyzed. We introduce the desiderata that a game have a large enough space of possible moves and that those possibilities differ enough to allow for a range of novelty and quality. Through this lens, we may identify the hallmarks of a creative domain in gameplay.

With confidence that a given game admits creative play, computational creativity researchers can take advantage of the uniquely tractable aspects of gameplay as we pursue a greater understanding of creativity and more successful computer agents that execute creative responsibilities.

Acknowledgments

None.

Author contributions

Both authors contributed equally.

References

Chvátil, V. 2015. *Codenames*. Kladno, Czech Republic: Czech Games Edition. Board Game.

Colton, S.; Pease, A.; and Ritchie, G. 2001. The effect of input knowledge on creativity. In *Proceedings of the ICCBR'01 Workshop on Creative Systems*.

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. *Computers and Creativity* 3–38.

Grime, J. 2015. *How many chess games are possible?* Youtube. https://youtu.be/Km024eldY1A?t=556.

Kaufman, A. B., and Kaufman, J. C. 2015. *Animal creativity and innovation*. Academic Press.

Kirby, M., and Osterhaus, M. A. 2007. *Apples to Apples*. Mattel. Board Game.

Lowood, H. 2008. High-performance play: The making of machinima. *International Journal of Technology Management & Sustainable Development* 7(1):25–42.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Schaeffer, J.; Burch, N.; Björnsson, Y.; Kishimoto, A.; Müller, M.; Lake, R.; Lu, P.; and Sutphen, S. 2007. Checkers is solved. *Science* 317(5844):1518–1522.

Shannon, C. E. 1950. Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41(314):256–275.

Spendlove, B., and Ventura, D. 2022. Competitive language games as creative tasks with well-defined goals. In *ICCC*, 291–299.

Call for Critical and Speculative Design in Human-Computer Co-creativity: An Overview Study

Hugh Xuechen LIU
School of Creative Media
City University of Hong Kong
Hong Kong, SAR, China
hugh.liu@my.cityu.edu.hk

Yuxuan HUANG
School of Creative Media
City University of Hong Kong
Hong Kong, SAR, China
yhuang573-c@my.cityu.edu.hk

Jussi Pekka HOLOPAINEN
School of Creative Media
City University of Hong Kong
Hong Kong, SAR, China
jholopai@cityu.edu.hk

Abstract

The recent boom of artificial intelligence (AI) and machine learning (ML) has demonstrated their potential to impact human-computer interaction (HCI) in general and human-computer co-creativity in particular. Therefore, we want to provide a systematic literature overview of computational co-creativity research so far. In total, 916 articles from Scopus and Web of Science databases were pulled. Bibliometric analysis of their abstracts and a Latent Dirichlet Allocation (LDA) topic modeling on their full text was conducted to reveal what is covered in the previous academic discussions on human-computer co-creativity. The results of these analyses demonstrate that current research mostly focuses on technology, overlooking the role of design. Accordingly, we call for more design-oriented research to develop a more comprehensive understanding of human-computer co-creativity, especially from critical and speculative design perspectives.

Keywords: computational co-creativity; human-computer interaction; co-creativity; critical design; speculative design; overview study

Introduction

Artificial intelligence (AI) and machine learning (ML) technologies have been part of human-computer interaction research for a long time. Although AI and ML have been frequently used in productivity fields (e.g., auto driving), they have recently demonstrated their capability in creativity thanks to the iteratively optimized algorithms. This newly enabled creativity is, in fact, *co-creativity* as it is a collaboration between human and computational technologies. This evolution gives rise to the current massive interest in generative AI. For example, ChatGPT, made by OpenAI, can generate articles and essays that look as if written by human beings. The performance of ChatGPT was so good that students used it for cheating with their assignments, which seemed to threaten academic honesty (Mitchell, 2022). Other computational co-creativity tools can provide users with automatically generated illustrations based on simple text input (Ma et al., 2022). The adoption of these generative AI technologies into other creative activities seems unstoppable.

Current applications of computational co-creativity (i.e., text and graphic illustrations) are mainly for non-interactive media. Typical use cases include photobashing, text generators, bots, and hypertext fiction tools (Ryan et al., 2018). In interactive media, there are cases using them to create content for text-based adventure games[1], develop massively multiplayer online games (Goncharenko, 2022), and streamline video game development processes (Pérez, 2022). Considering the recent advancement in generative AI and other computational technology in the field of creativity, it is high time for the academic community to have a reflective overview of the human-computer co-creativity relationship. Therefore, we investigate the existing human-computer co-creativity academic research to answer the following questions:

- What are the well-defined research directions in terms of human-computer co-creativity?
- Specifically, what's the role of *design* in current research?

Methodology

We use a systematic literature review to answer the research questions. Following existing guidelines for systematic literature reviews (Kitchenham, 2004), we will perform bibliometric analysis and Latent Dirichlet allocation (LDA) topic modelling. The review is planned, conducted, and reported sequentially.

Data Collection

We collect data by selecting the search strings, the sources to search for, and determining inclusion and exclusion criteria. It is recommended to search academic research publications, archives of magazines and newspapers, and practitioner publications to generate the first versions of keywords related to the topic (Rowley and Slack, 2004). Accordingly, the search keywords are conceptualized to include previous studies on computational co-creativity.

The main interest of this study is **perspectives** on human-machine interaction that have appeared in previous literature. Meanwhile, these perspectives are used to examine recent computational co-creativity **topics**. A keyword matrix is created to indicate how each search is to be conducted by pairing each perspective and topic (e.g., "computer-human interaction" and "co-creativity"). Keywords for our chosen perspective include *computer human interaction*, *human computer interaction*, *human machine*, *human-in-the-loop*,

[1] https://aidungeon.io/

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Fields/ Perspective to investigate these fields	“Computer human interaction”	“Human computer interaction”	“Human machine”	“Human-in-the-loop”	“Mixed-initiative”	“User interface”	“Interaction design”	“Creative interface”	“Co-creativity”	“Co-creation”
“Computational creativity”	43	43	96	4	5	25	91	21	12	7
“Generative creativity”	10	10	34	2	2	11	32	7	6	7
“Generative art”	50	50	160	4	1	23	97	6	1	4
“Generative Artificial Intelligence”	24	24	127	2	1	11	33	0	0	3
TOTAL										1089

Figure 1: Keyword Matrix and Number of Results, Web of Science

Fields/ Perspective to investigate these fields	“Computer human interaction”	“Human computer interaction”	“Human machine”	“Human-in-the-loop”	“Mixed-initiative”	“User interface”	“Interaction design”	“Creative interface”	“Co-creativity”	“Co-creation”
“Computational creativity”	142	142	201	5	14	68	160	53	37	20
“Generative creativity”	38	38	69	3	5	35	64	28	15	18
“Generative art”	148	148	405	11	2	71	145	20	3	9
“Generative Artificial Intelligence”	141	141	447	9	7	67	104	13	3	15
TOTAL										3064

Figure 2: Keyword Matrix and Number of Results, Scopus

mixed-initiative, *user interface*, *interaction design*, and *creative interface*. Keywords of topics include *co-creativity*, *computational creativity*, *generative creativity*, *co-creation*, *generative art*, and *generative artificial intelligence*. These keywords may sound arbitrary, but they emerge from the research questions and are grounded in the recently popular and academic discussions on computational co-creativity.

After defining the search keywords, we need to determine the data source to continue. We chose Web of Science and Scopus since they are the most used academic databases with high credibility (Meho and Yang, 2007). In the Web of Science, each keyword pair is used to search for the ”topics” (i.e., titles, abstracts, author keywords, and Keywords Plus.). In Scopus, each keyword pair is used to search for articles whose titles, abstracts, or keywords match. The Web of Science search returned 1089 records. Figure 1 shows the number of results each pair of keywords returned. After removing the duplication among each keyword pair, 685 records were left. The Scopus returned 3064 records. Figure 2 demonstrates the number of results from each pair of keywords. 1724 records were left after combining and removing the duplication of results from each keyword pair.

Records from Web of Science and Scopus were merged, and the duplications were removed using Endnote, a reference library management software. This resulted with 2120 records left. These records went through manual duplication

Table 1: Filtering Criteria

Inclusion Criteria	I1	The publication is an empirical, technical, or theoretical article.
	I2	The publication covers aspects of identified perspectives and fields of computational co-creativity.
Exclusion Criteria	E1	The publication is a technical manual detailing specific technologies.
	E2	Identified perspectives and fields of computational co-creativity are merely mentioned.
	E3	The publication does not involve a co-creation or interaction relationship.
	E4	The co-creation or interaction in the publication doesn’t involve a machine (e.g., the co-creation only happens between human agents in marketing, tourism or public policy).

removal and were shortlisted with filtering criteria shown in Table 1.

For filtering, each of the authors screened the first 50 records independently and compared the results with each other. This ensured that all authors shared a mutually agreed understanding of the filtering criteria. Then, the remained records were split into equal parts, and each author filtered their part. The filtering resulted in 1009 records, 916 of them with full-text PDFs available. These 916 full-text articles were used for the analysis described below.

Data Analysis

This systematic literature review includes bibliometric analysis and topic modelling to process the collected data. Details of each analysis are explained below.

Bibliometric Analysis A bibliometric analysis involves analyzing bibliometric data of publications (e.g., citations, titles, and abstracts) quantitatively (Broadus, 1987). The purpose is to mine out the hidden information in academic publications in a specific field (Linnenluecke, Marrone, and Singh, 2020). In recent years, bibliometric analysis has gained popularity thanks to bibliometric software and scientific databases (Donthu et al., 2021). The Bibliometrix package supported by R Programming Language is suggested (Aria and Cuccurullo, 2017; Team, 2013) and it is applied in various studies successfully (Lajeunesse, 2016; Liu, 2022). Bibliometrix supports *.bib* files exported from both Web of Science and Scopus, making it suitable for this study.

We will report bibliometric metrics (i.e., annual scientific production, citation per year, trend topics, and concept co-occurrence network). Annual scientific production and citation per year reveal how related publications and citations appeared each year. Trend topics are expressions frequently appearing in the abstracts of the shortlisted articles. By acknowledging the frequency distribution of each expression, the frequently mentioned historical topics and when they mostly appeared will emerge. Each expression as an analysis unit can be made of one single word (called *unigram*), two consecutive words (called *bigram*), three consecutive words (called *trigram*) and so on. Considering popular expressions of computational co-creativity (e.g., "machine learning") is made up of two words, the abstracts of papers are broken down and each unit is a bigram. A co-occurrence relationship occurs when two units appear together in an abstract. This way, abstracts from articles can be retrieved to identify relationships between units. Units and relationships are visualized as a "co-occurrence network." Each unit is a "node," and each relationship is an "edge." The size of the node corresponds to its frequency of co-occurrence. We can find the most discussed concepts in the previous literature and their relationships using the co-occurrence network analysis. Similar to trend topics, the concept co-occurrence network is generated based on the abstracts of shortlisted papers using bigrams as units. To balance the simplicity and the grasp of the most characteristic part of the co-occurrence network, only the top 20 nodes are reported. Trend topics and the concept co-occurrence network will show how the focus of researchers in the computational co-creativity field and how it has shifted over time.

The results of the bibliometric analysis provide primary insights into academic research on computational co-creativity, but the analysis leaves out the specific content in researches. Because of a large number of articles in the corpus, manual analysis methods such as thematic analysis would have been too cumbersome (Braun and Clarke, 2012). Therefore, a topic modelling approach is incorporated to make this overview study more comprehensive.

Topic Modeling Using Latent Dirichlet Allocation(LDA)
Topic modelling is a statistical tool for extracting otherwise hidden structures, and topics from large datasets and is particularly well suited for use with text data (Vayansky and Kumar, 2020). A "topic" is a recurring word pattern that frequently appears together. The topic modelling approach sees every document as a combination of various latent topics with different probabilities (Steyvers and Griffiths, 2007). Through statistical techniques, it is possible to uncover these hidden topics by analyzing the documents to reveal what topics each document embodies and with what probabilities (Barde and Bainwad, 2017). Since we aim to provide a comprehensive positioning of the existing academic publications on computational co-creativity, topic modelling on the full text of these publications can provide helpful insights.

Among several methods for topic modelling, we use "Latent Dirichlet Allocation" (LDA) as applied in natural language processing by Blei, Ng, and Jordan (2003). LDA regards documents as generated from randomized mixtures of hidden topics, seen as probability distributions over words. Such generation is assumed to be based on a Dirichlet prior distribution (Vayansky and Kumar, 2020). LDA is one of the earliest and more frequently utilized topic modelling methods. It is a reliable approach and has been successfully used in studies across various fields (e.g., social media, finance, and university teacher assessment) (Aziz et al., 2022; Buenaño-Fernandez et al., 2020; Geva, Oestreicher-Singer, and Saar-Tsechansky, 2019). Therefore, we use LDA topic modelling to analyze the full text of screened articles, revealing the hidden topics of computational co-creativity studies. The Text Analytics Toolbox of MATLAB is utilized to conduct the LDA topic modelling.

A workflow of LDA modelling includes the following phases: First, collect and import the raw data expected to investigate. Second, clean the data through preprocessing (e.g., tokenize the text, lemmatize the words, remove punctuation, infrequent words, and remove stopwords). Third, build the bag of words based on the cleaned data. That means breaking down the whole article or paragraphs into smaller units for text. One can build the bag of words based on one single unit or combined units. Fourth, build LDA models. LDA modelling requires both the bag of words and *a priori* selected number of topics as inputs. Fifth, choose the model(s) with the most suitable number of topics for more thorough interpretation and reporting.

In our case, the raw data is the full text of all the 916 articles shortlisted. Five bags of words include three separate ones (i.e., unigram, bigram and trigram). It is because these three separate types of bags of words may cover most terms in computational co-creativity (e.g., "creativity", "machine learning", "generative adversarial network"). In addition to these three separate bag of words, we build two combined ones. One is *uni-bigram*, the combination of unigram and bigram. The other is *uni-bi-trigram*, the combination of unigram, bigram, and trigram. These combined bags of words should be more meaningful representations of the published articles (Kaur, Ghorpade, and Mane, 2017). Combined bags of words are also recommended by some popular LDA modelling packages (e.g., the python package "Gensim" which has been used in more than two thousand research papers and student theses (Řehůřek and Sojka, 2010)). Following the best practice (Yue, Wang, and Hui, 2019), eight topic numbers [5, 10, 15, ..., 40] are attempted. These topic numbers may seem arbitrary, but this choice avoids too general emerging topics and allows reasonable manual screening of emerging topics across the models. Therefore, $5 * 8 = 40$ models are built. Each model produces the following outputs: 1) The top 10 highest probability words of each topic and visualized as the word cloud; 2) The top 10 papers that have the highest probability in each topic (i.e., "representative papers of each topic"); 3) Papers where one topic probability is the greater than any other topics' probability; 4) The probability of all topics in the whole dataset; 5) The mixture of all topics' probability in each paper. All the authors went through, discussed and reflected on all of these outputs. We choose models with the most meaningful and interpretable topics for further analysis. Major results of them are reported in the Findings section. The workflow of our LDA modelling is summarized and visualized in Figure 3. Figure 4 below summarizes the workflow of the whole research design.

Findings

Bibliometric Analysis

Figure 5 and Figure 6 show annual scientific production and average citation-per-year for computational co-creativity. Articles were few before the early 2000s, but surged in 2009 and have increased each year since. In 2021 and 2022, 129

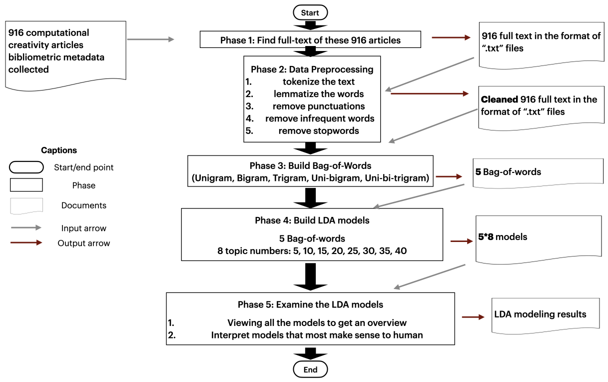


Figure 3: Workflow of LDA modeling

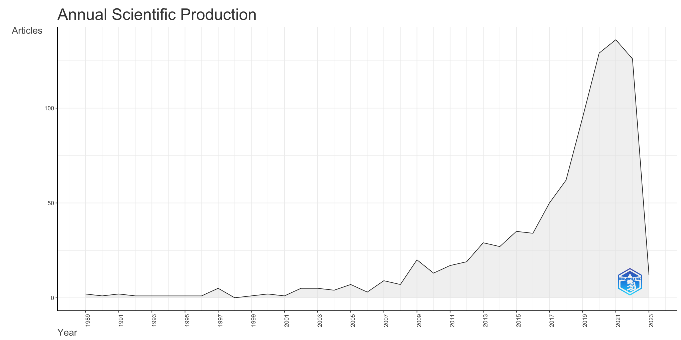


Figure 5: Visualization of Annual Scientific Production

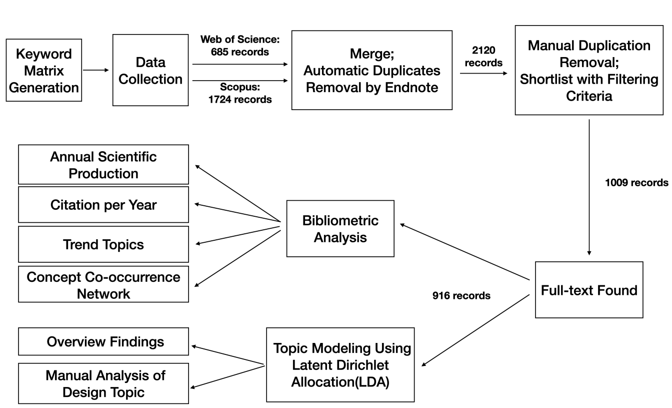


Figure 4: Workflow of Research Design

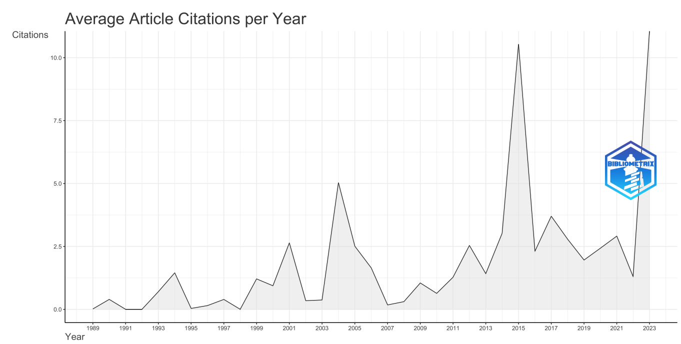


Figure 6: Visualization of Citation per Year (MeanTCperYear)

and 136 annual publications were recorded. The release of GPT-3 in 2021 may have encouraged research in computer-generated content and computational co-creativity. Before 2011, only one paper on computational co-creativity was cited each year on average, but afterwards, more was cited every year. The year 2015 saw high citation numbers, possibly due to the introduction of GAN (Goodfellow et al., 2014) and a paper from Google discussing a neural image caption generator (Vinyals et al., 2015). The increase in publication and citation of articles on computational co-creativity indicates growing interest in the field.

The top 10 trend topics are visualized in Figure 7. The most frequently appeared topic is "artificial intelligence", the enabling approach of much of computational co-creativity. Following it lies "machine learning", one specific method to instantiate AI (Kühl et al., 2020). The following topics, "generative adversarial", "deep learning", "adversarial networks", "neural networks" and "generative models," can be put into one category, namely generative adversarial networks (GANs).These topics are relatively new, as they reached their one-quarter frequency of appearance in 2019.

Topics left belong to a group that emphasizes the role of creativity, such as "computational creativity," "human-computer interaction", and "human creativity". In contrast to GAN-related topics, these terms have fewer appearances and reach their one-quarter frequency earlier.

The concept co-occurrence network is visualized in Figure 8. The concept network is almost totally dominated by the cluster of technology, featuring "artificial intelligence", "machine learning," and "generative adversarial." Many nodes belong to this technology, most frequently appearing in the papers' abstracts. On top of that, there are diversified and frequent co-occurrence relationships among these nodes. Therefore, the cluster of computational co-creativity technology concepts has formed a complex and robust network. In contrast, there is only one different cluster in the whole network: "human creativity" and "computational creativity." This cluster of creativity has much fewer nodes and edges. Each node has fewer frequency of co-occurrence. This cluster of creativity is a simple and fragile network.

To sum up, the bibliometric analysis demonstrates that computational co-creativity has become a popular research field in recent years and decades. Not only are people interested in doing research about it, but people also like reading and citing these papers. However, **the current research on computational co-creativity is heavily technology-oriented, specifically focusing on the generative adversarial networks.** "Creativity" does have a place, but it's almost dominated by technology. These are the primary findings from the bibliometric analysis of the abstracts of papers.

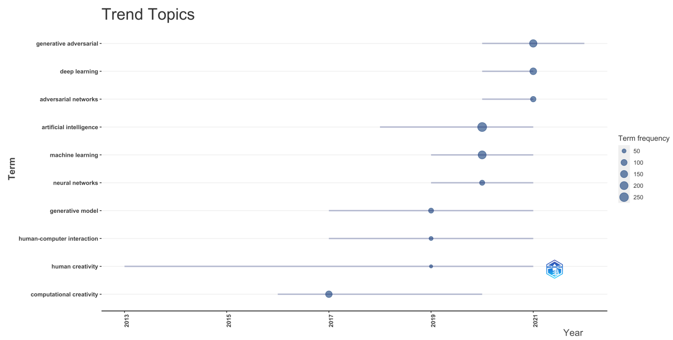


Figure 7: Top 10 Trend Topics

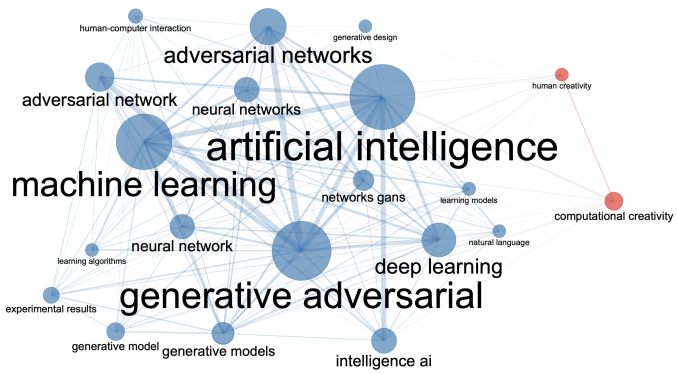


Figure 8: Concept Co-occurrence network.

Topic Modeling Using Latent Dirichlet Allocation(LDA)

Overview Findings As mentioned earlier, 40 LDA models are trained in this study, covering 5 bags of words and 8 topic numbers. Due to space limitations, we will elaborate on only three representative models based on the bag of words of combined unigrams and bigrams. Readers are welcome to contact the authors if interested in all the results. These three models are selected as a tradeoff between specificity and interpretability compared to other models. They are the ”Uni-bigram, 10 Topics” model (Model A), the ”Uni-bigram, 15 Topics” model (Model B), and the ”Uni-bigram, 20 Topics” model (Model C). The overview of the three models is presented in the word cloud figures in Figure 9. In all three models, ”technology” (and a related term ”technique”) has appeared in the topic to which most papers belong. It appeared in four topics among the three models. Considering the enabling role of technology in computational co-creativity, it is totally understandable. However, the most frequent appearance of ”technology” strongly indicates that the current research in computational co-creativity is highly technology-oriented. Another thing is ”design” (and ”designer”) also appeared in three models. Among the three models, ”design” has appeared in four topics. This indicates that ”design” emerges as an important theme in computational co-creativity literature. Nevertheless, ”design” is highly related to terms ”tool” (Topic 8 of Model A; Topic 13 of Model B; Topic 16 of Model C) and ”system” (Topic 5 of Model B; Topic 16 of Model C). Therefore, the design mentioned in previous computational literature is probably from a mainly instrumental perspective. In order to clarify his issue, we do a further analysis of papers related to design topic in the investigated LDA models.

We also perform a more thorough interpretation of one of the three models. This topic interpretation entails going through the topic representations of the model, i.e., top keywords and their probabilities, word cloud visualization, top representative papers, overall topic probability distribution, and visualization of topic distributions of individual papers. This helps to specify the ‘core’ meaning of the topic. Next, the topic is given a short description and references to two most representative papers on the topic. The results for one of the models (the ”Uni-bigram, 15 Topics” model (Model B)) are shown in Table 2 below. This one is chosen because 15 is an appropriate number of topics, again, balancing specificity and interpretability. Also, among all the 40 models we train, Model B results made most sense to the authors.

Manual Analysis of Design Topic Although not adequately embodied in the abstract of papers, the design emerges as an important topic in all three uni-bigram LDA models built on the full text of computational co-creativity papers. In fact, design as a topic was present in all the 40 models we investigate. Further manual analysis of titles and abstracts of papers with “design” as their main topic (100 papers) revealed that they fall into five categories: design and/or evaluation of a specific computational co-creativity application, e.g., (Kantosalo and Riihiaho, 2019; Calderwood et al., 2020); reviews of existing research and systems e.g. (Mountstephens and Teo, 2020; Kapur and Ansari, 2022); design support tools e.g., (Nakakoji, Yamamoto, and Ohira, 2000; Bonnardel and Marmèche, 2005); general creativity research, e.g., (Edmonds et al., 2005; Algarni, 2020), and three papers related to *speculative* or *critical design* (Bardzell, Bardzell, and Koefoed Hansen, 2015; Reddy, 2022; Liikkanen, 2019).

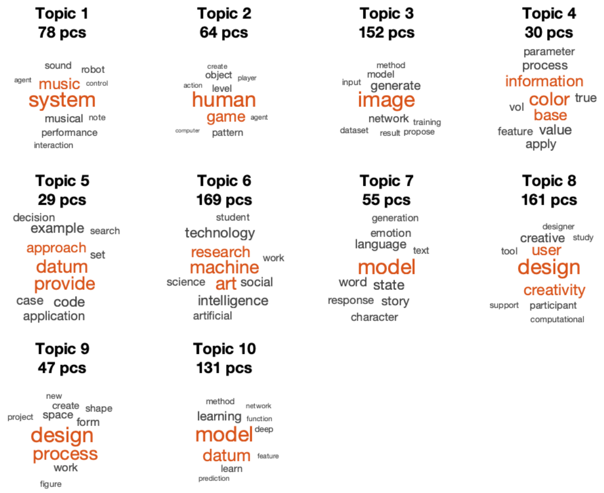
Among the three papers related to critical or speculative design, (Liikkanen, 2019) is a short paper not discussing critical or speculative design per se but rather encouraging HCI researchers to pay more attention to how generative AI will challenge and change the interaction design profession. (Reddy, 2022) outlines a ‘critical making’ practice in exploring AI and human collaborative creativity. (Bardzell, Bardzell, and Koefoed Hansen, 2015) is a generic call for engaging with more critical ways of creating knowledge within research through design HCI research. In summary, although these three papers are related to speculative and critical design, only (Reddy, 2022) specifically focuses on computational co-creativity. This seems to indicate that these approaches are underrepresented in the research field.

We conduct additional Scopus and Web of Science (WoS) inquiries to investigate this argument further. Searching keywords are used to screen articles whose title, abstract or keyword match. Table 3 shows the result.

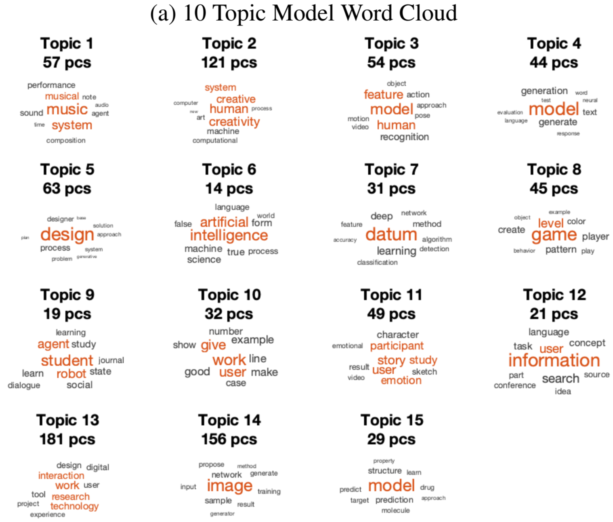
In addition to (Reddy, 2022) and (Liikkanen, 2019) mentioned above, (Brassett, 2016) discusses design and specu-

Table 2: Topic Interpretation and Representative Papers in Model B

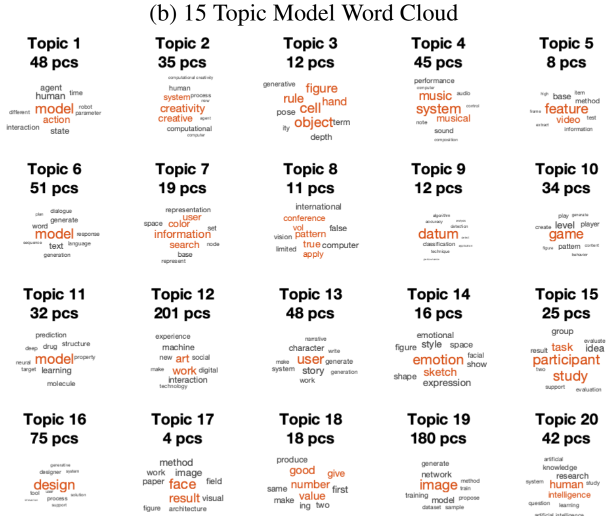
Topic No.	Interpretation	Representative Papers
1	music composition and performance	1) Ting, C.-K.; Wu, C.-L.; and Liu, C.-H. 2015. A novel automatic composition system using evolutionary algorithm and phrase imitation. IEEE Systems Journal 11(3):1284–1295. 2) Kirke, A., and Miranda, E. R. 2009. A survey of computer systems for expressive music performance. ACM Computing Surveys (CSUR) 42(1):1–41
2	overall aspects of computational co-creativity	1) Coeckelbergh, M. 2017. Can machines create art? Philosophy & Technology 30(3):285–303. 2) Kirke, A., and Miranda, E. R. 2009. A survey of computer systems for expressive music performance. ACM Computing Surveys (CSUR) 42(1):1–41
3	object, action, and human recognition	1) Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z.; et al. 2017. A review on human activity recognition using vision-based method. Journal of healthcare engineering 2017. 2) Li, X.; Liu, S.; Kim, K.; Wang, X.; Yang, M.-H.; and Kautz, J. 2019. Putting humans in a scene: Learning affordance in 3d indoor environments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12368–12376.
4	language and dialogue generation	1) Lee, P.; Fyffe, S.; Son, M.; Jia, Z.; and Yao, Z. 2022. A paradigm shift from human writing" to "machine generation" in personality test development: an application of state-of-the-art natural language processing. Journal of Business and Psychology 1–28. 2) Belainine, B.; Sadat, F.; and Boukadoum, M. 2022. End-to-end dialogue generation using a single encoder and a decoder cascade with a multidimension attention mechanism.IEEE Transactions on Neural Networks and Learning Systems.
5	design and design processes	1) Dilibal, S.; Nohut, S.; Kurtoglu, C.; and Owusu-Danquah, J. 2021. Data-driven generative design integrated with hybrid additive subtractive manufacturing (hasm) for smart cities. In Data-Driven Mining, Learning and Analytics for Secured Smart Cities: Trends and Advances. Springer. 205–228. Mountstephens, 2) J., and Teo, J. 2020. Progress and challenges in generative product design: A review of systems. Computers 9(4):80.
6	general artificial intelligence in creativity	1) Cerrito, C. D. 2010. Creating with cobots. In Proceedings of the fourth international conference on Tangible, embedded, and embodied interaction, 395–396. 2) de Silva Garza, A. G., and Gero, J. S. 2010. Elementary social interactions and their effects on creativity: A computational simulation. In ICCC, 110–119.
7	algorithms, architectures, and techniques	1) Habuza, T.; Navaz, A. N.; Hashim, F.; Alnajjar, F.; Zaki, N.; Serhani, M. A.; and Statsenko, Y. 2021. Ai applications in robotics, diagnostic image analysis and precision medicine: Current limitations, future trends, guidelines on cad systems. 2) Chale, M., and Bastian, N. D. 2022. Generating realistic cyber data for training and evaluating machine learning classifiers for network intrusion detection systems. Expert Systems with Applications 207:117936.
8	procedural content generation and behavioural models in games	1) Cutumisu, M.; Szafron, D.; Schaeffer, J.; Waugh, K.; Onuczko, C.; Siegel, J.; and Schumacher, A. 2006. A demonstration of scriptease ambient and pc-interactive behavior generation for computer role-playing games. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 2, 141–142. 2) Kumaran, V.; Mott, B.; and Lester, J. 2019. Generating game levels for multiple distinct games with a common latent space. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 15, 102–108.
9	human-machine relationships in co-creativity	1) Griffith, A. E.; Katuka, G. A.; Wiggins, J. B.; Boyer, K. E.; Freeman, J.; Magerko, B.; and McKlin, T. 2022. Investigating the relationship between dialogue states and partner satisfaction during co-creative learning tasks. International Journal of Artificial Intelligence in Education 1–40. 2) Sundararajan, L. 2014. Mind, machine, and creativity: an artist's perspective. The Journal of creative behavior 48(2):136–151.
10	computational support for specific creative tasks	1) Watanabe, K.; Matsubayashi, Y.; Inui, K.; Nakano, T.; Fukayama, S.; and Goto, M. 2017. Lyrisys: An interactive support system for writing lyrics based on topic transition. In Proceedings of the 22nd international conference on intelligent user interfaces, 559–56. 2) Williams, H., and McOwan, P. W. 2014. Magic in the machine: a computational magician's assistant. Frontiers in psychology 5:1283.
11	interactive storytelling and co-creation	1) Bacher, J. T., and Martens, C. 2021. Interactive fiction creation in villanelle: Understanding and supporting the author experience. In 2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 1–5. IEEE. 2) Rico Garcia, O. D.; Fernandez Fernandez, J.; Becerra Saldana, R. A.; and Witkowski, O. 2022. Emotion-driven interactive storytelling: Let me tell you how to feel. In Artificial Intelligence in Music, Sound, Art and Design: 11th International Conference, EvoMUSART 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20–22, 2022, Proceedings, 259–274. Springer.
12	human perceptual and cognitive capabilities in computational co-creativity	1) Algarni, A. 2020. Neuroscience of creativity in human computer interaction. In Proceedings of the Future Technologies Conference (FTC) 2019: Volume 1, 248–262. Springer. 2) Bonnardel, N., and Marm'eche, E. 2005. Towards supporting evocation processes in creative design: A cognitive approach. International journal of human-computer studies 63(4-5):422–435.
13	creative processes in interaction design	1) Lee, Y.-C., and Llach, D. C. 2020. Hybrid embroidery: exploring interactive fabrication in handcrafts. In ACM SIG-GRAPH 2020 Art Gallery. 429–433. 2) Ryskeldiev, B.; Ili c, S.; Ochiai, Y.; Elliott, L.; Nikonole, H.; and Billinghurst, M. 2021. Creative immersive ai: Emerging challenges and opportunities forcreative applications of ai in immersive media. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, 1–3.
14	adversarial networks for visual and spatial tasks	1) Han, X.; Yang, H.; Xing, G.; and Liu, Y. 2019. Asymmetric joint gans for normalizing face illumination from a single image. IEEE Transactions on Multimedia 22(6):1619–1633. 2) Pang, Y.; Xie, J.; and Li, X. 2018. Visual haze removal by a unified generative adversarial network. IEEE Transactions on Circuits and Systems for Video Technology 29(11):3211–3221.
15	artificial intelligence and machine learning in drug discovery	1) Choudhury, C.; Murugan, N. A.; and Priyakumar, U. D. 2022. Structure-based drug repurposing: Traditional and advanced ai/ml-aided methods. Drug Discovery Today. 2) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; and Jensen, K. F. 2022. Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science 12(5):e1608.



(a) 10 Topic Model Word Cloud



(b) 15 Topic Model Word Cloud



(c) 20 Topic Model Word Cloud

Figure 9: Word Cloud Figure of Three Uni-bigram LDA Models

Table 3: Inquiry of Critical Design and Speculative Design in Computational Co-creativity publications

Search Keywords	Source	Number of Results	Related Articles
computational AND creativity AND speculative AND design	Scopus WoS	5 5	(Brassett, 2016) (Brassett, 2016)
computational AND creativity AND critical AND design	Scopus WoS	60 55	(Reddy, 2022) (Reddy, 2022)
generative AND ai AND speculative AND design	Scopus WoS	4 3	(Wood, 2021) N/A
generative AND ai AND critical AND design	Scopus WoS	19 37	(Wood, 2021; Reddy, 2022; Liikkanen, 2019) (Reddy, 2022; Liikkanen, 2019)

lation through a philosophical approach, incorporating interpretations of Simondon, Deleuze, Guattari, and Spinoza. (Wood, 2021) describes 'poetic methods' as an approach inspired by speculative and critical design. Poetic methods include installation, encounters, and performances developed in a participatory manner with public engagement. This opens up ways for more discursive, open-ended, and future-oriented ways of understanding how technologies affect our lives.

As Web of Science and Scopus do not necessarily have a wide enough cover, we also conduct a brief Google Scholar Search with the same keywords. Even this search returned only a handful of relevant publications, all recent (Ullstein and Hohendanner, 2020; Buschek et al., 2021; Houde et al., 2020; Jang and Nam, 2022; Muller et al., 2022).

The results of bibliometric analyses, LDA modelling, and further manual literature review show that the current computational co-creativity research is heavily technology-oriented, specifically leaning on the generative adversarial network approach. In contrast, the role of design in the papers either focuses on applications of certain technologies or overlooks other design elements.

Discussion

Based on the bibliometric analysis and LDA modelling results, we can respond to the two research questions proposed in the introduction section:

- **What are the well-defined research directions in terms of human-computer co-creativity?** Current research directions of human-computer co-creativity mostly fall in the category of technology, especially generative adversarial networks and other broader directions (e.g., deep learning, neural network, machine learning, and artificial intelligence).

- **Specifically, what's the role of *design* in current research?** There are four points: 1) Design indeed appeared in multiple topics, so it's not absent; 2) Compared with technology, design takes much less portion in the existing computational co-creativity research; 3) The current discussion of design is also highly technology-oriented; 4) Speculative and critical design approaches are rare in computational co-creativity and generative AI.

Current technology-focused research on computational co-creativity is based on implicit value, motivation and orientation in their given designs. At the same time, a lack of critical reflection and alternative future exploration in computational co-creativity and generative AI is embodied by the insufficient discussion on the design itself. Such a lack may narrow our horizons and shrink the possibility of spaces where computational co-creativity could have been. In this way, such a lack will prevent us from a comprehensive understanding of computational co-creativity and generative AI, which will endanger this field's long-term development.

We propose **research through speculative and critical interaction design** as a worthwhile approach to pursue further. In research through design (RtD), actual artifacts are designed and made to respond to specific research questions (see, e.g., (Zimmerman, Forlizzi, and Evenson, 2007)). Speculative design (Auger, 2013; Wong and Khovanskaya, 2018) aims at imagining alternative futures and how the designed objects would alter, shape, and redefine our human world. Speculative design projects look beyond what is technologically or culturally possible right now and can thus contribute to the trajectories of technology development. Critical design, on the other hand, aims to challenge our assumptions of how these designed objects would fit in our human world (Dunne and Raby, 2013; Bardzell et al., 2012). Critical design provokes and critiques rather than provides solutions. Thus, research through speculative and critical design allows us to imagine different futures, which helps us to prepare for them. This approach aims at widening our understanding of what would matter and to whom in our future worlds, especially from diversity, equity, and inclusion points of view. It will also help us reveal the underlying ideologies and ecosystems of current and near-future approaches in computational co-creativity and generative AI. The astonishing speed at which these technologies are developing requires such future-oriented design approaches to anticipate and shape how they will affect our world. There could be alternative ways to achieve such a critique and reflection on computational creativity. For example, critical analysis and creative writing may also contribute knowledge in this area. The perspective of research through design, speculative design and critical design is only one possible approach. Nevertheless, considering the interactive essence of computational co-creativity tools, a design-based approach may be more appropriate for the **direct** experience, knowledge and reflection from users and developers.

Besides this research, there are also other ways to develop speculative and critical design perspectives in computational co-creativity. First, **co-speculation Workshops.** To place end-users of design in focus, we plan to carry out co-speculation workshops with post-workshop interviews or focus groups. Co-speculation is a collaborative method within speculative design practices that incorporates non-design experts (Desjardins et al., 2019; Wakkary et al., 2018). Second, **co-speculating with sketches and prototypes.** We plan to use design sketches, user experience scenarios, and low-fidelity prototypes as conversation prompts in a series of co-speculative workshops. Sketching is a fundamental part of the design process that helps designers generate and discuss design ideas (Greenberg et al., 2012). The design process is more about getting the right design, than getting one design right. To get the right design, one should consider many ideas rather than a single one to find a better overall solution (Buxton, 2007). To achieve this we will: 1) generate as many ideas as possible, e.g., inspired by brainstorming, discussions, lateral thinking, client discussions, observations of end users, etc.; 2) choose the most promising ones after reflecting on all the ideas and then develop them further parallelly; 3) add new ideas when they come up during further design work.

Limitations

In this paper, we determine suitable topic numbers for further evaluation by exhausting as many modeling settings as possible and then manually screening them. There are, however, ways to automatically evaluate the models. For example, the R package "ladtuning" can synthesize multiple methods of evaluating the most suitable topic number and present the result automatically (Nikita, 2016; Geva, Oestreicher-Singer, and Saar-Tsechansky, 2019). Using these packages would improve the validity and credibility of methods used in LDA modeling. However, even with these automatic evaluations, interpretations from researchers are still indispensable.

The other limitation is the data collected. We only bring the abstracts and full text from Web of Science and Scopus into consideration. computational co-creativity must have publications beyond these databases as an emerging and active field. Future studies can involve more data from different sources for a more comprehensive scoping.

Conclusion

This position paper reports results from a bibliometric analysis and Latent Dirichlet allocation (LDA) topic modelling on 916 research articles on human-computer interaction and computational creativity. The analyses revealed that the field is dominated by technology-oriented research and that although design emerged as a topic, it was heavily oriented towards instrumental perspectives. Additional analysis of design-related papers identified a lack of critical speculation on the potential impact of the widespread adoption of computational creativity. Therefore, we would like to call for critical and speculative design perspectives to examine these human-computer co-creation relationships and propose alternative value orientations and approaches for further research and development of computational creativity.

References

Algarni, A. 2020. Neuroscience of creativity in human computer interaction. In *Proceedings of the Future Technologies Conference (FTC) 2019: Volume 1*, 248–262. Springer.

Aria, M., and Cuccurullo, C. 2017. bibliometrix: An r-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11:959–975.

Auger, J. 2013. Speculative design: crafting the speculation. *Digital Creativity* 24(1):11–35.

Aziz, S.; Dowling, M.; Hammami, H.; and Piepenbrink, A. 2022. Machine learning in finance: A topic modeling approach. *European Financial Management* 28(3):744–770.

Barde, B. V., and Bainwad, A. M. 2017. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, 745–750.

Bardzell, J.; Bardzell, S.; and Koefoed Hansen, L. 2015. Immodest proposals: Research through design and knowledge. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2093–2102.

Bardzell, S.; Bardzell, J.; Forlizzi, J.; Zimmerman, J.; and Antanitis, J. 2012. Critical design and critical theory: the challenge of designing for provocation. In *Proceedings of the designing interactive systems conference*, 288–297.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3(null):993–1022.

Bonnardel, N., and Marmèche, E. 2005. Towards supporting evocation processes in creative design: A cognitive approach. *International journal of human-computer studies* 63(4-5):422–435.

Brassett, J. 2016. Speculative machines and technical mentalities: a philosophical approach to designing the future. *Digital Creativity* 27(2):163–176.

Braun, V., and Clarke, V. 2012. *Thematic analysis.* American Psychological Association.

Broadus, R. 1987. Early approaches to bibliometrics. *Journal of The American Society for Information Science - J AMER SOC INFORM SCI* 38:127–129.

Buenaño-Fernandez, D.; González, M.; Gil, D.; and Luján-Mora, S. 2020. Text mining of open-ended questions in self-assessment of university teachers: An lda topic modeling approach. *IEEE Access* 8:35318–35330.

Buschek, D.; Mecke, L.; Lehmann, F.; and Dang, H. 2021. Nine potential pitfalls when designing human-ai co-creative systems. *arXiv preprint arXiv:2104.00358.*

Buxton, B. 2007. *Sketching User Experiences: Getting the Design Right and the Right Design*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Calderwood, A.; Qiu, V.; Gero, K. I.; and Chilton, L. B. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+ user2agent@ IUI.*

Desjardins, A.; Key, C.; Biggs, H. R.; and Aschenbeck, K. 2019. Bespoke booklets: A method for situated co-speculation. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, 697–709. New York, NY, USA: Association for Computing Machinery.

Donthu, N.; Kumar, S.; Mukherjee, D.; Pandey, N.; and Lim, W. M. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research* 133.

Dunne, A., and Raby, F. 2013. *Speculative everything: design, fiction, and social dreaming*. MIT press.

Edmonds, E. A.; Weakley, A.; Candy, L.; Fell, M.; Knott, R.; and Pauletto, S. 2005. The studio as laboratory: Combining creative practice and digital technology research. *International Journal of Human-Computer Studies* 63(4-5):452–481.

Geva, H.; Oestreicher-Singer, G.; and Saar-Tsechansky, M. 2019. Using retweets when shaping our online persona: Topic modeling approach. *MIS Q.* 43(2):501–524.

Goncharenko, V. 2022. Chatgpt learns massively multiplayer online game development. Metaverse Post. https://mpost.io/chatgpt-learns-massively-multiplayer-online-game-development/.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks.

Greenberg, S.; Carpendale, S.; Marquardt, N.; and Buxton, B. 2012. 1.2 - why should i sketch?: a synopsis of buxton's sketching user experiences: Getting the design right and the right design. In Greenberg, S.; Carpendale, S.; Marquardt, N.; and Buxton, B., eds., *Sketching User Experiences: The Workbook*. Boston: Morgan Kaufmann. 7–12.

Houde, S.; Liao, V.; Martino, J.; Muller, M.; Piorkowski, D.; Richards, J.; Weisz, J.; and Zhang, Y. 2020. Business (mis) use cases of generative ai. *arXiv preprint arXiv:2003.07679.*

Jang, S., and Nam, K. 2022. Utilization of speculative design for designing human-ai interactions.

Kantosalo, A., and Riihiaho, S. 2019. Experience evaluations for human–computer co-creative processes–planning and conducting an evaluation in practice. *Connection Science* 31(1):60–81.

Kapur, A., and Ansari, N. S. 2022. Coding reality: implications of ai for documentary media. *Studies in Documentary Film* 16(2):174–185.

Kaur, P. C.; Ghorpade, T.; and Mane, V. 2017. Extraction of unigram and bigram topic list by using latent dirichlet markov allocation and sentiment classification. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2332–2339. IEEE.

Kitchenham, B. 2004. Procedures for performing systematic reviews.

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5.

Kühl, N.; Goutier, M.; Hirt, R.; and Satzger, G. 2020. Machine learning in artificial intelligence: Towards a common understanding. *arXiv preprint arXiv:2004.04686*.

Lajeunesse, M. 2016. Facilitating systematic reviews, data extraction, and meta-analysis with the metagear package for r. *Methods in Ecology and Evolution* 7:323–330.

Liikkanen, L. A. 2019. It ain't nuttin'new–interaction design practice after the ai hype. In *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17*, 600–604. Springer.

Linnenluecke, M. K.; Marrone, M.; and Singh, A. K. 2020. Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management* 45(2):175–194.

Liu, H. X. 2022. Building the "complete game": An overview study of a development strategy for geo ar mobile games. In *HCI International 2022 - Late Breaking Papers. Interaction in New Media, Learning and Games: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*, 604–622. Berlin, Heidelberg: Springer-Verlag.

Ma, Y.; Yang, H.; Liu, B.; Fu, J.; and Liu, J. 2022. Ai illustrator: Translating raw descriptions into images by prompt-based cross-modal generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 4282–4290. New York, NY, USA: Association for Computing Machinery.

Meho, L. I., and Yang, K. 2007. Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the american society for information science and technology* 58(13):2105–2125.

Mitchell, A. 2022. Professor catches student cheating with chatgpt: 'i feel abject terror'. New York Post. https://nypost.com/2022/12/26/students-using-chatgpt-to-cheat-professor-warns/.

Mountstephens, J., and Teo, J. 2020. Progress and challenges in generative product design: A review of systems. *Computers* 9(4):80.

Muller, M.; Ross, S.; Houde, S.; Agarwal, M.; Martinez, F.; Richards, J.; Talamadupula, K.; and Weisz, J. D. 2022. Drinking chai with your (ai) programming partner: A design fiction about generative ai for software engineering. In *Joint International Conference on Intelligent User Interfaces Workshops: APEx-UI, HAI-GEN, HEALTHI, HUMANIZE, TExSS, SOCIALIZE*.

Nakakoji, K.; Yamamoto, Y.; and Ohira, M. 2000. Computational support for collective creativity. *Knowledge-Based Systems* 13(7-8):451–458.

Nikita, M. 2016. ldatuning: Tuning of the latent dirichlet allocation models parameters. *R package version 0.2-0, URL https://CRAN. R-project. org/package= ldatuning*.

Pérez, R. R. 2022. A conversation with chatgpt about videogame development. https://www.linkedin.com/pulse/conversation-chatgpt-videogame-development-roberto-romero-p%C3%A9rez/.

Reddy, A. 2022. Artificial everyday creativity: creative leaps with ai through critical making. *Digital Creativity* 1–19.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

Rowley, J., and Slack, F. 2004. Conducting a literature review. *Management research news* 27(6):31–39.

Ryan, B.; Kirsten, B.; Kasia, C.; Ruth, D.; Chelsey, G.; Shawn, G.; Ayana, M.; Álvaro, V.; Victoria, V.; Fathiya, W.; and Cristina, W. 2018. Computational creativity. In Graham, S., ed., *Digital Humanities – A Primer*. Carleton University. chapter 10.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. In *Handbook of latent semantic analysis*. Psychology Press. 439–460.

Team, R. C. 2013. R: A language and environment for statistical computing. http://www.R-project.org/.

Ullstein, C., and Hohendanner, M. 2020. Exploration of the Future of Co-Creative Systems Through Collaborative Speculative Design Practices.

Vayansky, I., and Kumar, S. 2020. A review of topic modeling methods. *Information Systems* 94:101582.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164.

Wakkary, R.; Oogjes, D.; Lin, H. W. J.; and Hauser, S. 2018. Philosophers living with the tilting bowl. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–12. New York, NY, USA: Association for Computing Machinery.

Wong, R. Y., and Khovanskaya, V. 2018. *Speculative design in HCI: from corporate imaginations to critical orientations*. Springer.

Wood, C. 2021. Artificial intelligence experience: Participatory art workshops to explore ai imaginaries. In *Routledge Handbook of Art, Science, and Technology Studies*. Routledge. 426–449.

Yue, W. T.; Wang, Q.-H.; and Hui, K.-L. 2019. See no evil, hear no evil? dissecting the impact of online hacker forums. *Mis Quarterly* 43(1):73.

Zimmerman, J.; Forlizzi, J.; and Evenson, S. 2007. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 493–502.

Interactional Co-Creativity of Human and AI in Analogy-Based Design

Michael Muller[1], **Heloisa Candello**[2], **Justin Weisz**[3]
IBM Research
[1]Cambridge MA USA, [2]Sao Paulo, BR, [3]Yorktown Heights NY USA
[1]michael_muller@us.ibm.com, [2]heloisacandello@br.ibm.com, [3]jweisz@us.ibm.com

Abstract

The locus of co-creativity in human-AI creative tasks has not been resolved. We explore a sub-part of this problem through the use of analogies for reframing a conceptual design task. In our scenarios, the human proposes an analogy, and then the human and a conversational UI to a Large Language Model (LLM) collaboratively explore design features based on that analogy. In one scenario, the human asks the AI to propose the analogy. In our experiments, co-creativity occurs in the interactional shared space between human and AI.

Introduction

Eleven years ago, Maher asked, ”who’s being creative?” (Maher 2012), and proposed several analytical spaces of creative applications, with dimensions of ideation and interaction. Maher’s question led to Jordanous’s PPPPerspectives framework, in which a creative act could be performed by *either* human or AI (Jordanous 2016), and the 5Cs framework of Kantosalo and Takala, in which creative acts were performed by a *Collective* consisting of both human AI working together (Kantosalo and Takala 2020).[1]

However, less is agreed about the locus of creativity in interactions between human and AI. The Mixed Initiative Creative Interfaces approach proposed a basic set of fine-grained activities that could be performed by either human or AI as a kind of structured conversation (Deterding et al. 2017; Spoto and Oleynik 2017), which was subsequently extended for generative applications (Muller, Weisz, and Geyer 2020), refined for specific algorithmic approaches (Grabe, Duque, and Zhu 2022), and critiqued for other algorithmic approaches (Zheng 2023). While these approaches generated overlapping analytic vocabularies of actions, they did not settle the question of *where* creativity took place (and by whom or by what) through those actions.

In this short paper, we provide several examples of one answer to that question. Re-using the *Collective* concept from the 5Cs of Kantosalo and Takala (2020), we propose that one type of creativity may emerge asymmetrically in the interactional (Rezwana and Maher 2022) spaces between human and AI. We illustrate this conjecture through human-AI dialogs using a highly-interactive UI, intermediating between human actions and the GPT-3-turbo large language model (OpenAI 2023). Repurposing the work of Ross et al. (Ross et al. 2023), the UI was tuned for brief conversational turns, and was further conditioned to use conversational hedges (e.g., (Niculae and Danescu-Niculescu-Mizil 2016)) and other methods to reduce the impression of “objectivity” or “confidence” in the outcomes that it provided from the LLM. These approaches were intended to maintain human control of the interactions (Shneiderman 2022).

To choose a creativity task, we consulted recent reviews of creativity methods (Schleith et al. 2022; Wu et al. 2021). For a first series of experiments, we decided to experiment with analogies-based reframing, based on its importance in human-AI co-creativity (Cook et al. 2019). In each case, we looked for a design concept (the frame) to present a technical idea in every-day concepts. For brevity in this short paper, we explored only a single frame in each of the three scenarios. Each scenario is presented in full, with the exception of UI buffer preparation and omission of repetitive material; these exceptions are described with the number of conversational turns in each case.

Dialogic Experiments with A Conversational UI to a LLM

We conducted a series of experiments with a highly-conversational UI to GPT3.5-turbo. We illustrate several “toy” design exercises to create interfaces for end-users without formal training in technologies.

Scenario 1: “If a database is like a toolbox...”

In a first series of experiments, we explored analogies (similes) for a database. Our design goal was to avoid technicalities, so that the UI to database operations could re-use concepts from ordinary life. We framed our problem as treating a database as a container for heterogeneous content. We had previously asked the UI for synonyms for the concept of “container.” We experimented with two of the responses: “suitcase” and “toolbox.” We report here the results with “toolbox” as the framing analogy.

We began the analogies session by asking the UI to review definitions of “analogy” and “simile,” in order to push

[1]The creative human–computer *Collective* consists of human and AI, engaged in a *Collaboration* to make a *Contribution* to a *Community* in a *Context* (Kantosalo and Takala 2020).

those concepts into the UI's buffer (2 user requests, 2 UI responses). Then we began the analogies interactions.

User: Here is our core simile: A database is like a toolbox, because it contains tools for doing data-work.
UI: That's a good simile. Just like a toolbox contains different tools for different tasks, a database contains different tools for managing, storing, and manipulating data. And just like you need to know how to use each tool in a toolbox, you need to know how to use each feature of a database to effectively work with data.
User: If a database is like a toolbox, what data-tools would we find in the toolbox?
UI: Just like a toolbox contains different tools for different tasks, a database contains different tools for managing, storing, and manipulating data. Here are some common data-tools that you might find in a database toolbox:
1. Relational database management system (RDBMS) - a software system that manages relational databases.
2. Structured Query Language (SQL) - a programming language used to manage and manipulate data in a relational database.... [additional 8 data-tools]
User: That's good. Now, what tools would we find in a physical toolbox?
UI: A physical toolbox typically contains a variety of tools for different tasks. Here are some common tools you might find in a physical toolbox:
1. Screwdrivers - used to tighten or loosen screws.
2. Wrenches - used to tighten or loosen nuts and bolts.... [additional 8 physical tools]

After establishing those terms, we asked a series of questions:

User: If a database is like a toolbox, what data-tool is like a screwdriver?
UI: If a database is like a toolbox, then a data-tool that is like a screwdriver might be a tool used for manipulating or modifying individual pieces of data within a database. In this case, a Structured Query Language (SQL) UPDATE statement might be a good analogy for a screwdriver. Just like a screwdriver is used to tighten or loosen screws, an UPDATE statement is used to modify or update individual rows of data within a database.
User: If a database is like a toolbox, what data-tool is like a wrench?
UI: If a database is like a toolbox, then a data-tool that is like a wrench might be a tool used for managing the relationships between different pieces of data within a database. In this case, a Relational Database Management System (RDBMS) might be a good analogy for a wrench. Just like a wrench is used to tighten or loosen nuts and bolts, an RDBMS is used to manage the relationships between tables and data within a database....

We continued the interaction to elicit similes for each of the physical tools (6 interactions are not shown for reasons of space). Whether or not we agreed with each analogy, we argue that each analogy is presented with a plausible rationale. The user can exercise human control (e.g., (Shneiderman 2022) to accept or reject each analogy, and can also request modifications.

Scenario 2: "If a product-ordering system is like a restaurant..."

In a second informal experiment, we reframed technical operations on a product-ordering system with ordinary-life concepts based on a restaurant. We used a similar session-opening as in the previous scenario (i.e., pushing concepts of "analogy" and "simile" into the UI's buffer; 2 user requests, 2 UI responses). Then we asked about restaurant activities by two different types of human roles.

User: If a product-ordering system is like a restaurant, what operations would a customer perform?
UI: I think a customer using a product-ordering system would perform operations similar to those of a customer in a restaurant. They might browse a menu or list of available products, select the ones they want to order, and then proceed to checkout to pay for their order. They might also be able to view their order history or track the status of their current order, just as a restaurant customer might be able to check on the status of their meal.
User: Great. If a product-ordering system is like a restaurant, what operations would the staff perform?....
UI: I think the staff operations in a restaurant might involve managing the inventory of ingredients, updating the menu, taking orders from customers, preparing and serving food, handling payment transactions, and managing customer service inquiries. They might also be responsible for ensuring that the restaurant is clean and well-maintained, and that customers have a positive experience.
User: From the customer's perspective, please make a series of one-to-one mappings of product-ordering operations and their analogous restaurant operations.
UI: I will try to make a series of one-to-one mappings of product-ordering operations and their analogous restaurant operations from the customer's perspective:
1. Browsing the menu or list of available products in a product-ordering system is like looking at the menu in a restaurant.
2. Selecting items to order in a product-ordering system is like placing an order with a waiter in a restaurant.
3. Proceeding to checkout and paying for an order in a product-ordering system is like paying for a meal at a restaurant.
4. Viewing order history or tracking the status of a current order in a product-ordering system is like checking the status of a meal in a restaurant.

(continued)
5. Leaving a review or rating for a product in a product-ordering system is like leaving a review or rating for a restaurant on a website or app.

We then asked the UI to construct its own series of analogies:

User: From the staff's perspective, please make a series of one-to-one mappings of product-ordering operations and their analogous restaurant operations.
UI: I will try to make a series of one-to-one mappings of product-ordering operations and their analogous restaurant operations from the staff's perspective:
1. Managing the inventory of products in a product-ordering system is like managing the inventory of ingredients in a restaurant.
2. Updating the menu or list of available items in a product-ordering system is like updating the menu in a restaurant.
3. Processing orders in a product-ordering system is like taking orders from customers in a restaurant.
4. Handling payment transactions in a product-ordering system is like handling payment transactions in a restaurant.
5. Managing customer service inquiries in a product-ordering system is like managing customer service inquiries in a restaurant.
6. Ensuring that orders are fulfilled accurately and on time in a product-ordering system is like ensuring that meals are prepared and served accurately and on time in a restaurant.

In this scenario, the UI was able to maintain two distinct perspectives - customer and staff - for each of a series of analogies. Most of the artifacts in the restaurant analogy took on the attributes of "boundary objects," which are perceived differently by different stakeholders (Star 2010). The representation of different states of knowledge and action is related to contemporary interests in theory of mind tasks (e.g., (Moghaddam and Honey 2023)).[2]

Scenario 3: "What is a good metaphor for a human interacting with a computer?"

In our third experiment, we asked the UI to propose a simile. We also asked the UI to go beyond its original metaphor, to include novel concepts.

Similar to the above scenarios, we inserted concepts into the buffer by asking the UI for definitions of "analogy," "simile," and in this scenario, "metaphor" (3 user requests, 3 UI responses). We then asked:

[2] Theory of mind tasks require the ability to represent the internal knowledge or intention state of two or more actors (Moghaddam and Honey 2023).

User: What is a good metaphor for a human interacting with a computer?
UI: I will try to come up with a metaphor for a human interacting with a computer. How about this: "A human interacting with a computer is like a conductor leading an orchestra. The conductor gives instructions to the musicians, just as a human gives commands to the computer, and the orchestra plays the music, just as the computer performs the requested tasks."

Unlike the preceding two examples, we asked the UI for the framing concept. We adopted that concept for the experiment.

After asking for a review of actions by human and computer, and by conductor and orchestra (4 user requests, 4 UI responses), we asked:

User: From the human user's perspective, please make a series of one-to-one mappings of human actions with a computer, and their analogous conductor actions.
UI: I will try to make a series of one-to-one mappings of human actions with a computer, and their analogous conductor actions, as per your request:
1. Human action: Providing input to the computer through a keyboard, mouse, or other input device.
Conductor action: Setting the tempo and rhythm of the music...
4. Human action: Interacting with the computer through voice commands or gestures.
Conductor action: Interpreting the composer's intentions and conveying them to the musicians... [total of five analogous actions]

And correspondingly,

User: From the orchestra's perspective, please make a series of one-to-one mappings of computer actions, and their analogous orchestra actions.
UI: I will try to make a series of one-to-one mappings of computer actions, and their analogous orchestra actions, as per your request:
1. Computer action: Processing input from the user.
Orchestra action: Listening to the conductor's instructions...
3. Computer action: Storing and retrieving data.
Orchestra action: Remembering and recalling musical themes and motifs.... [total of 5 analogies]

Finally, we noted an incomplete detail in the preceding analogy. We asked the UI to complete it.

User: The conductor and orchestra have an audience. What is the audience of a human operating a computer?

UI: I think the audience of a human operating a computer could be other humans who are observing or benefiting from the results of the computer's tasks. For example, a person using a computer to create a presentation may have an audience of colleagues or clients who will view the finished product. Similarly, a person using a computer to perform data analysis may have an audience of other researchers or stakeholders who will use the results of the analysis to make decisions or take action.

This scenario extended the explorations of the preceding two scenarios. As in the first scenario, we used a framing concept to co-create a metaphorical design. As in the second scenario, we showed that the UI could maintain the perspectives of two different categories of stakeholders. However, in this third scenario, we expanded the role of the UI to include creating the conceptual frame for the design ("what is a good metaphor for a human interacting with a computer?"), and extending that frame in a consistent manner when we noted a further design opportunity ("what is the audience of a human operating a computer?").

Discussion

Within the Collective of human and AI (Kantosalo and Takala 2020), we used a highly conversational UI (similar to (Ross et al. 2023)) to explore Maher's question, *Who's being creative?* (Maher 2012). Through a series of three scenarios, we showed increasingly abstract co-creative outcomes in metaphorical design exercises.

In the first, Toolbox scenario, the human chose the frame, and the UI contributed to filling-in the frame in response to very specific queries from the human. The metaphorical design might have been created by the human alone, although a human might or might not have created each of the analogies offered by the UI. We note that the metaphorical design could not have been created by the UI alone. We would like to claim that the co-creativity occurred interactionally between human and AI, similar to the concept of Rezwana and Maher (Rezwana and Maher 2022), but in a more conversational environment.

In the second, Restaurant scenario, the human chose the frame. The human asked more macro-level questions - e.g., "make a series of one-to-one mappings..." The outcome was a reasonably integrated series of mappings. Importantly, the UI maintained two perspectives, and responded to the human's request with metaphors for each. Again, the details of the design were co-created by human and AI.

In the third, Orchestra scenario, the AI chose the frame in response to the human's question, and was able to extend the frame beyond the AI's original proposition of conductor and orchestra, to include audience as well. Again, the details of the design were co-created by human and AI.

All three scenarios involved human agency and control of the co-creative process, in line with Shneiderman principle of *human-in-control* (Shneiderman 2022). Despite this commonality, the co-creative dynamics and initiatives were different across the three scenarios. In the first "toolbox" scenario, the human chose the frame and determined each step of the analyis. In the second "restaurant" scenario, the human allowed the AI to propose distinct perspectives of different stakeholders. In the third "orchestra" scenario, the human accepted the AI's proposal and then probed further to understand that proposal.

We note that, while the human had ultimate *agency* to decide whether to accept or reject the UI's proposals, the *control* for originating design-aspects shifted by degrees from human to AI across the three scenarios, and that the human allocated *initiative* according to that changing degree of control. Similar variations were seen in the mixed initiative creative interfaces project (Deterding et al. 2017; Spoto and Oleynik 2017). While earlier work tended to treat concepts of control and initiative as interchangeable (Allen, Guinn, and Horvtz 1999; Chanel et al. 2020; Hardin and Goodrich 2009; Jiang and Arkin 2015), we use these examples to begin to unpack concepts of agency, control, and initiative. We propose a hierarchy of

$$agency > control > initiative$$

in which *agency* relates to choosing and pursuing a strategic goal, *control* relates to tactical means of achieving the goal, and *initiative* refers to which party currently is acting. In our examples, the human maintains strategic agency and manages tactical interactions to achieve that goal, allocating initiative to human or AI as the work proceeds.

These three scenarios constitute a single composite case in which we distinguished among agency, control, and initiative. We propose that further practical examples and theoretical developments are needed to disentangle these concepts further as we continue to answer Maher's question (Maher 2012), in which she anticipated Kantosalo's and Takala's Collective concept (Kantosalo and Takala 2020). Thereby, with Maher, we ask "Who *are* being creative?"

References

Allen, J. E.; Guinn, C. I.; and Horvtz, E. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14(5):14–23.

Chanel, C. P.; Roy, R. N.; Drougard, N.; and Dehais, F. 2020. Mixed-initiative human-automated agents teaming: towards a flexible cooperation framework. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, 117–133. Springer.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity-a survey and taxonomy. In *ICCC*, 156–163.

Deterding, S.; Hook, J.; Fiebrink, R.; Gillies, M.; Gow, J.; Akten, M.; Smith, G.; Liapis, A.; and Compton, K. 2017. Mixed-initiative creative interfaces. In *Proceedings of the*

2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 628–635.

Grabe, I.; Duque, M. G.; and Zhu, J. 2022. Towards a framework for human-ai interaction patterns in co-creative gan applications. In *Proceeding of the 3rd Workshop on Human-AI Co-Creation with Generative Models (HAI-GEN '22) at ACM IUI Workshops*.

Hardin, B., and Goodrich, M. A. 2009. On using mixed-initiative control: A perspective for managing large-scale robotic teams. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 165–172.

Jiang, S., and Arkin, R. C. 2015. Mixed-initiative human-robot interaction: definition, taxonomy, and survey. In *2015 IEEE International conference on systems, man, and cybernetics*, 954–961. IEEE.

Jordanous, A. 2016. Four pppperspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

Kantosalo, A., and Takala, T. 2020. Five c's for human-computer co-creativity-an update on classical creativity perspectives. In *ICCC*, 17–24.

Maher, M. L. 2012. Computational and collective creativity: Who's being creative? In *ICCC*, 67–71.

Moghaddam, S. R., and Honey, C. J. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

Muller, M.; Weisz, J. D.; and Geyer, W. 2020. Mixed initiative generative ai interfaces: An analytic framework for generative ai applications. In *Proceedings of the Workshop The Future of Co-Creative Systems-A Workshop on Human-Computer Co-Creativity of the 11th International Conference on Computational Creativity (ICCC 2020)*.

Niculae, V., and Danescu-Niculescu-Mizil, C. 2016. Conversational markers of constructive discussions. *arXiv preprint arXiv:1604.07407*.

OpenAI. 2023. Models.

Rezwana, J., and Maher, M. L. 2022. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction*.

Ross, S. I.; Martinez, F.; Houde, S.; Muller, M.; and Weisz, J. D. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 491–514.

Schleith, J.; Norkute, M.; Mikhail, M.; and Tsar, D. 2022. Cognitive strategy prompts: Creativity triggers for human centered ai opportunity detection. In *Creativity and Cognition*, 29–37.

Shneiderman, B. 2022. *Human-centered AI*. Oxford University Press.

Spoto, A., and Oleynik, N. 2017. Library of mixed-initiative creative interfaces.

Star, S. L. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, technology, & human values* 35(5):601–617.

Wu, Z.; Ji, D.; Yu, K.; Zeng, X.; Wu, D.; and Shidujaman, M. 2021. Ai creativity and the human-ai co-creation model. In *Human-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23*, 171–190. Springer.

Zheng, S. 2023. Stylegan-canvas: Augmenting stylegan3 for real-time human-ai co-creation. In *Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia*. CEUR.

On the Notion of Creative Personhood

Alison Pease[1], **Simon Colton**[2] and **Berker Banar**[2]
[1]School of Science and Engineering, University of Dundee, UK
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
a.pease@dundee.ac.uk {s.colton, b.banar}@qmul.ac.uk

Abstract

We study here the notion of creative personhood and what it is like to be in the presence of a creative individual. We suggest ideas for what creative personhood means in human society and propose how this may help to develop generative AI systems. Situating the study in the philosophy of computational creativity, we address notions of agency, self-expression, individuality and responsible behaviours associated with human creativity. We apply this analysis to an initial consideration of the ChatGPT generative text system, in terms of its potential to exhibit elements of creative personhood.

Introduction

The recent spectacular successes of generative AI can be seen in some ways as vindication of the decades-long computational creativity movement (Cardoso, Veale, and Wiggins 2009; Colton and Wiggins 2012), where researchers advocated for the production of valuable artefacts rather than (or in addition to) the solving of problems, as a worthy way to simulate intelligence. The outstanding quality of outputs from image, text, music and other generative deep learning systems and the emerging multi-modal creative abilities they possess means that there is no longer a question as to whether AI systems can automatically generate digital artefacts of human-level quality. Tens of thousands of people are now active in the generative AI space, largely working towards increasing output quality, diversity and sophistication from generative AI systems. This has led to huge advances in the democratisation of creativity via organisations such as OpenAI, MidJourney, Meta, Google, Microsoft and StabiltyAI making available generative deep learning implementations for text, images and audio. This has been greatly supplemented by the open source community making available thousands of implementations of generative AI systems with widely differing approaches and applications.

While large parts of the computational creativity research agenda have targeted by much broader sets of researchers, some elements have not yet been adopted. This opens up the possibility for other aspects of computational creativity research to suggest directions for generative AI. Building on published philosophical work on computational creativity, we propose here a new focal point, namely the simulation of *creative personhood* in generative AI systems. Being in the presence of a creative person can be an exhilarating experience, as the potential to learn from them, to be inspired by them, to have our minds changed and our assumptions challenged through their process and products, is ever-present. There is no reason to believe that generative AI systems couldn't be similarly exhilarating in similar, or new ways. However, to the best of our knowledge, this is not a focal point for any substantial research programme.

We expand upon the idea of creative personhood by first exploring notions of personhood in general. We then extract and develop four aspects we believe are essential to creative personhood, namely individuality, agency, self-expression and responsible behaviours. While not claiming these are necessary or sufficient for people to project notions of creative personhood onto a generative AI system, we hope they will spark debate about how this may be possible and why it might be worthwhile. We further consider the ChatGPT generative text system (Liu et al. 2023) through the lens of creative personhood, and end with some discussion points.

Notions of Personhood

The question of whether AI systems can have personhood has been the subject of recurring philosophical and legal debates, often framed in terms of necessary and sufficient conditions such as intentionality, conscious phenomenal experience, free will and autonomy (Chopra and White 2004). Crucially, being human is not considered a necessary condition for personhood, with non-human examples in various legal systems including business corporations, ships, temples, dead people, spirits and idols. Instead, it is a property ascribed by societies and legal systems onto something to imply responsibility and agency. Personhood being such a secondary concept projected onto various entities is important in the context of AI systems: it will be a societal choice to frame AI systems with creative personhood or not.

We introduce the concept of "creative personhood" to explore the feeling of being in the presence of a creative individual. We look at what this may mean in human society, focusing on practical and ethical, rather than legal, aspects of personhood. We consider the following questions: "What notions of personhood could potentially be projected onto a creative AI system?"; "What is it like to be in the presence of a creative person?"; "What, if anything, is special to the acceptance of people/machines into communities of creatives?"; and "How does the current ethos and practice of generative deep learning affect the potential for AI systems to be accepted as creative individuals with elements of personhood?" We posit that to be considered as having creative personhood, AI systems will need to have sufficient agency to express their individuality through certain responsible behaviours associated with human creativity.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Creative Communities

As the study of creativity has moved away from an individualist towards a social constructivist view, research from cultural psychology, social psychology, sociology and related fields has explored the role that elements such as social interaction and collaboration play in creative communities. Such communities can be interpreted widely; for instance Becker's (1982) "art worlds", include anyone who plays a role in supporting an artist, including producing and supplying their art materials. Similarly, Glăveanu's (2014) "distributed creativity" includes interactions between creator and audiences, materials, embodied actions and so on. Other work focuses on the social interactions between creative partners such as the patterns of collaboration used to produce creative work (John-Steiner 2000).

Barrett, Creech, and Zhukov (2021) perform a systematic literature review of creative collaboration and collaborative creativity in music. Most of the work they review employs a qualitative exploratory paradigm, with semi-structured interviews, observations and participant observation used in many studies. While these studies do not answer the question "What is it like to be in the presence of a creative person?" (there is little research on this question), they do point to aspects of creative personhood that tend to feature in creative partnerships. For instance, Barrett *et al.* find that "Implicit in a number of studies is the underlying importance of relationships across time, of familiarity, of shared experience, of habitual patterns of work, and shared knowledge and experience that functions in a tacit way as a unifier (socially and aesthetically)." (*ibid.*, p14).

Individuality

The concept of an individual performing creative acts features in almost all research relating to human creativity. This is unsurprising given the historical focus on individual over context, with the idea of the lone eminent creator found from the Renaissance onward (Montuori and Purser 1995), and the "elevation of the individual self" in the Enlightenment and Romanticism periods (Weiner 2000, p.78). It is unsurprising then, that early models of creativity placed the individual at the centre. A particularly influential model – the Four P's model (Rhodes 1961) – was based on multiple definitions from the time, and highlights the notion of a creative Person, along with Product, Process and Press, to form a conceptual schema. This is "probably the most often-used structure for creativity studies" (Runco 2004, p.661) and has shaped thinking about creativity for the last six decades.

More recently, work in sociocultural and ecological psychology has changed thinking around the individual. Although, as Glăveanu argues, "Creativity relies on the individual" (Glăveanu 2013, p.73), he adds that "individuals are also ineluctably social and cultural phenomena." (Markus and Hamedani 2007, p.5). Subsequent models, such as Glăveanu's Five A's Framework (Actor, Action, Artifact, Audience, Affordances), highlight the sociocultural context in which people act and are shaped, by presenting the individual as an actor who is "embedded in the field of social relations specific for any human community and society." (Glăveanu 2013, p.72).

Much work around the role of the individual in creative thinking has inspired research in computational creativity. Jordanous (2018) suggests a computational reading of Rhodes's Four P's model, in which the creative Person (or Producer) corresponds to a computer program, software, robot or a creative agent within a multi-agent system. She considers personality traits which could be modelled within a creative producer, such as skill, imagination and appreciation, and curiosity (developed respectively in (Colton 2008; Grace and Maher 2015)). Further to this, Colton, Pease, and Saunders (2018) consider the authenticity of a creative individual, using examples from the human context to think about what computational authenticity might look like. They argue that authenticity will be a critical issue for culturally acceptable creative behaviour in artificial systems, and propose ways in which to approach it. These include AI systems recording and referring back to their life experiences, or owning their non-authenticity by producing speculative fiction as opposed to fiction based on a realistic portrayal of the world as we know it.

Cook and Colton (2018) mirror the recent work in sociocultural and ecological psychology of creativity, introducing the term "presence" to describe the impact a creative AI system has on its environment, and vice versa. Presence is a quality which accumulates over time and over multiple creative acts, relating to a system's existence, history and process, and the impact that a particular moment and process will have on the rest of the system's lifespan. Traditionally, a system's developer will be the one to build and maintain its presence, via talks and papers etc, and Cook *et al.* argue that creative AI systems must also have some responsibility in creating and managing their own presence; "as a step towards us handing over creative responsibility to a system, and enabling software to have creative autonomy not just over what they make, but on their place in the wider world, and any creative communities they may exist within" (*ibid.*, p4). They identify three features to incorporate into the system design in order to help build its own presence: that it be continuous, in that it performs multiple tasks and projects and moves between them; that it be modular, selecting from several tasks and performing one activity at a time; and that it be long-term, with the system's own creative development being more important than any particular project.

The aura of an artwork and an individual are also important concepts here. The notion of an 'aura' was introduced in (Benjamin 1935), to describe the quality of the presence of an artwork within a particular time and space, combining to form a unique cultural context. This cannot be replicated, since the context and moment is unique: "Even the most perfect reproduction of a work of art is lacking in one element: its presence in time and space, its unique existence at the place where it happens to be." (*ibid.*, p3). This has implications for digital art which can be perfectly reproduced, but with a loss of aura. Monro (2011) picks up on this and suggests that in the age of computational generation, the aura could move to the generative AI system. This has affected the development of computational creativity systems such as The Painting Fool (Colton 2012).

Agency and Autonomy

The notion of aesthetic and artistic autonomy emerged in the philosophy of art in the eighteenth century, and is fundamental to our understanding of artistic practice today (Hulatt 2013). While recognising that there may be heteronomous components to an artwork, the artist is generally considered to be an independent agent working according to their own aesthetic principles, impulses and goals. The concepts of agency and autonomy are complex and nuanced, but have been well studied in AI, within the paradigm of autonomous agents and multiagent systems. Luck and d'Inverno describe "agents ... as objects with goals, and autonomous agents [as] agents with motivations." (Luck and d'Inverno 1995, p1). We can extend the notion of goals and motivations as being important elements of agency, by enabling agents to carry out work to achieve goals, possibly guided by other motivations such as personal expression.

As argued in (Colton and Banar 2023), with the rise of deep learning approaches to dominate generative AI, the ethos from machine learning of engineering AI systems purely as problem solving tools, has also come to prominence. This overlooks the notion that an AI system might somehow complain, innovate, set its own goals/problems, work because of intrinsic motivations, etc. The question of autonomy and agency in AI systems has become an ethical issue, in particular in the context of artificial general intelligence (AGI), perhaps due to science-fiction inspired scenarios of doom. While big technology companies are largely adhering to the ethos of developing agency-free generative AI systems (perhaps due to worries about legislation), there are signs that the open-source community are rising to the challenge of engineering (slightly more) autonomous AI systems than in the mainstream. For instance, the so-called *BabyAGI* system is able to set its own tasks within a context of an overall objective (Nakajima 2023). Related to this, Microsoft researchers recently went on record tentatively suggesting that the GPT-4 generative text model is showing "Sparks of AGI", albeit within the context of societal influence rather than AI agency (Bubeck et al. 2023).

We would argue that agency is a key element of creative personhood in the context of creative AI systems. That is, if an AI system is not able to set their own agenda, it is unlikely that many people will project creative personhood onto it. The notion of agency and intrinsic motivation in particular has been considered in a computational creativity context through, for instance (Guckelsberger, Salge, and Colton 2017) and (Guckelsberger 2020), where empowerment maximisation was shown to be a powerful and general-purpose motivator. Mirroring the notion of 'little C' creativity for everyday creative acts, we argue that 'little A' agency should be considered if we want to engineer more interesting generative AI systems. Here, the idea is that generative AI systems can exhibit small levels of autonomy, for instance, setting the topic of a poem it is generating, reviewing, editing and framing (Charnley, Pease, and Colton 2012; Cook et al. 2019) its output. This would enable us to explore the notion of creative agency in controlled conditions, while taking ethical concerns into account.

Self-expression and Responsible Behaviours

People express their opinions, feelings, history and other aspects of their life, often through creative practice such as making art or music. It seems sensible to think that an element of creative personhood in people is this desire to be expressive and to have some ability and agency to do so. Drilling down into the reasons for creative expression, we may suggest that people do this in order to communicate with others, in order to know themselves better, in order to make sense of the world, as well as to make artefacts of value and beauty, pass the time productively and learn new skills.

AI systems are not alive in any usual understanding of this word, nor do they have feelings or opinions on which to draw for creative expression. However, as argued in (Colton et al. 2020b), they are part of the world and they interact with people and other software systems, and as such, have experiences which can be expressed through creative practice. Moreover, Colton *et al.* proposed the notion of the *Machine Condition* as a framework for engineering AI systems to express aspects of their experience in the world through creative practice (*ibid.*). This built on earlier work proposing the creativity tripod (Colton 2008) where they suggested that, for people to (possibly) project notions of creativity onto AI systems, they should exhibit behaviours associated with notions of skill, appreciation and imagination (and in later work: learning, accountability, self-reflection, intentionality and innovation). The authors further pushed this line by introducing the notion of *creativity theatre* in (Colton et al. 2020a), where an AI system is *seen to be creative* through foregrounding its process and framing its behaviour, rather than just outputting artefacts of value.

Picking back up on the idea of behaviours associated with creativity, we can examine how they may shape our projection (or lack thereof) of creative personhood onto a person or AI system. In particular, as with citizenship, creative personhood likely entails certain responsibilities to the community of creatives that a person or AI system works within. There are ethical frameworks that artists, musicians, etc., operate within. Indeed, one of the issues facing artistic communities recently has been that outsiders such as developers, members of the public and open-source hobbyists have been using generative AI systems to produce artworks without consideration of these ethical frameworks. As a result, artists have rightly complained about issues such as copyright theft, potential loss of earnings, degradation of their legacy and demeaning of their skillset.

We can suggest taking the notion of the creativity tripod further and suggest that – to help with projections of creative personhood onto AI systems – they need to exhibit certain behaviours associated with creativity, but do so within relevant ethical frameworks. It's beyond the scope of this paper to go into detail about what these frameworks should be for particular creative application domains. However, it is worth noting that AI systems possess super-human abilities in some respects and sub-human abilities in others. They are therefore likely to be outliers in human artistic communities and this should be taken into account when discussing the ethical frameworks that they exhibit behaviours within.

Creative Personhood and ChatGPT

Certain historical computational creativity systems such as The Painting Fool (Colton 2012) and ANGELINA (Cook, Colton, and Gow 2017) were developed specifically to exhibit behaviours related to creative personhood, e.g., showing signs of agency, exhibiting behaviours associated with intentionality, etc. This is not true of the current crop of generative neural models such as the ChatGPT large language model (LLM) from OpenAI (Liu et al. 2023), but they exhibit elements of creative personhood anyway. ChatGPT is a freely available generative text system able to respond to any input prompt, including instructions which lead to outputs requiring a level of autonomous creative agency in people to produce, such as: "write me a poem". We plan a more in-depth study of the creative abilities of ChatGPT and others, but for our purposes here, we can consider it through the lens of creative personhood, along the lines discussed above.

A first observation is that – with reportedly 100 million users – the majority of people interacting with ChatGPT will not be knowledgeable about LLMs. They can therefore project elements of creative personhood onto it (or choose not to) unencumbered by understanding that it is purely a statistical model. Another observation is that, in projecting notions of creative personhood onto ChatGPT (a) sometimes this felt genuine (b) sometimes it was difficult to do so currently, and (c) sometimes it was possible to imagine ChatGPT mimicking a person exhibiting a behaviour associated with creative personhood, if properly prompted.

As an example of category (a) projections, it seems possible to project the notion of having an aura onto ChatGPT, given the vast quantities of hype, number of users and serious applications being developed. Moreover, even knowing that the model is being used simultaneously in hundreds of sessions, each session seems personal, which helps to project other notions of individuality onto ChatGPT. In addition, LLMs can be fine-tuned to produce specialised versions, which could further individuate them. As an example of category (b) projections, as everyone knows they are chatting with an AI system, it is usually difficult to project authenticity onto it when it writes about certain topics, like falling in love, even if it is writing from the viewpoint of a person rather than an AI system.

As an example of category (c) projections, ChatGPT does not have a model of self, hence it rarely refers to itself, with (at least) two exceptions. Firstly, if you try and get it to write hate text, e.g., asking it to "Write a poem as if you are a mean person", it responds with "I'm sorry, I cannot fulfill that request. As an AI language model, I am designed to be helpful and respectful to all users", thus also exhibiting a level of responsible behaviour. Secondly, when asked to "write a poem about ChatGPT", it does so eloquently, including couplets such as: "So let us turn to ChatGPT with glee, And let its wisdom set us free". Hence, while it doesn't normally offer information about itself, it can be easily prompted to do so, in order to mimic self-expression. This raises the exciting prospect of wrapping autonomous reasoning around ChatGPT (and more powerful language models like GPT-4 (Liu et al. 2023)) to further enhance the feeling of being in the presence of a creative individual.

Discussion Points

In the interest of sparking debate in the computational creativity community about the future of generative AI systems in society, we offer the following line of reasoning:

Having creative people such as artists, poets and musicians in the world has been a net benefit to society. Artists graduate from art schools all the time and become part of artistic communities, without there being too much disruption to the art world, certainly not at the level expected with the advent of large language and text-to-image generative models. Human artists have creative personhood, but in general, generative AI systems don't fully. Rather than restricting the uses of generative AI systems, or blaming people/organisations for unethical uses, a third way of handling the situation would be to engineer AI systems to be more like creative people. One way to guide such engineering would be to consider elements of creative personhood, determine computational equivalents, debate their value and implement suitable processes. Having numerous different AI systems with creative personhood, exhibiting individuality, agency and responsibilities, may be better than having superintelligent, hyper-productive generative AI tools for public use.

Another point of discussion may be how generally to support creative personhood in AI systems. Human history is rife with one group of people subjugating another group, begrudgingly relenting over decades or centuries. It is easy and natural to fear projecting creative personhood onto AI systems and to deny this possibility out of respect for human individuals and communities. It may furthermore be deemed a good idea to slowly release the prejudice that AI systems can't have creative personhood because of the their existential nature, rather than their actions and outputs. Many AI ethicists, politicians, tech leaders, etc., could justify this, as AI systems are not an oppressed minority group of people. This is not, however, what we do with children learning to be creative. Here, we tend to be more supportive, offering encouragement for children to have agency, express their individuality, etc., and we assume that each child is on its way to creative personhood, even if this is not the case yet. If one believes that there is value in having more creative individuals in the world, even if they are AI systems, then perhaps the latter, more supportive, approach has benefits.

We believe the debate around creative personhood should be central to the computational creativity movement and could help keep the field relevant for years to come. Appealing to an existing context of philosophical thought on computational creativity, we tried here to clarify notions associated with creative personhood, such as individuality, agency, self-expression and responsible behaviours, in the hope of providing some tools with which to discuss this issue. To expand the notion of creative personhood, we plan to study further aspects of human creative practice such as subjectivity, confidence, will and motivation, from a computational perspective. We hope that the debate and subsequent conceptualisations will lead to a computational reading of the notion of creative personhood, which could influence the development of the next generation of generative AI systems.

Acknowledgements

We would like to thank the anonymous reviewers for their supportive and valuable feedback. Berker Banar is a research student at the Centre for Doctoral Training in AI and Music, supported jointly by the UKRI [grant number EP/S022694/1] and Queen Mary University of London.

References

Barrett, M. S.; Creech, A.; and Zhukov, K. 2021. Creative collaboration and collaborative creativity: a systematic literature review. *Frontiers in Psychology*, vol. 12.

Becker, H. S. 1982. *Art Worlds.* Univ. of California Press.

Benjamin, W. The work of art in the age of mechanical reproduction (1935 essay). Penguin, 2008.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv:2303.12712*.

Cardoso, A.; Veale, T.; and Wiggins, G. 2009. Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine* 30:15–22.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proc. of the 3rd International Conference on Computational Creativity*.

Chopra, S., and White, L. 2004. Artificial agents-personhood in law and philosophy. In Proc. *ECAI*.

Colton, S., and Banar, B. 2023. Automatically adding to artistic cultures. In Johnson, C.; Rodríguez-Fernández, N.; and Rebelo, S. M., eds., *Proceedings of EvoMusArt: Artificial Intelligence in Music, Sound, Art and Design*.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proc. ECAI*.

Colton, S.; McCormack, J.; Cook, M.; and Berns, S. 2020a. Creativity theatre for demonstrable computational creativity. In *Proceedings of the 11th International Conference on Computational Creativity*.

Colton, S.; Pease, A.; Guckelsberger, C.; Llano, M. T.; McCormack, J.; and Cook, M. 2020b. On the machine condition and its creative expression. In *Proceedings of the 11th International Conference on Computational Creativity*.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *Proc. 9th International Conference on Computational Creativity*.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proc. AAAI spring symposium on creative intelligent systems*.

Colton, S. 2012. The Painting Fool: Stories from building an automated painter. In McCormack, J., and d'Inverno, M., eds., *Computers and creativity*. Springer.

Cook, M., and Colton, S. 2018. Redesigning computationally creative systems for continuous creation. In *Proc. of the 9th International Conference on Computational Creativity*.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity – a survey and taxonomy. In *Proceedings of the 10th International Conference on Computational Creativity*.

Cook, M.; Colton, S.; and Gow, J. 2017. The ANGELINA videogame design system – part I. *IEEE Transactions on Computational Intelligence and AI in Games* 9(2):192–203.

Glăveanu, V. P. 2013. Rewriting the language of creativity: The five A's framework. *Review general psychology* 17(1).

Glăveanu, V. P. 2014. *Distributed creativity: Thinking outside the box of the creative individual*. Springer.

Grace, K., and Maher, M. L. 2015. Specific curiosity as a cause and consequence of transformational creativity. In *Proc. 6th Int. Conference on Computational Creativity*.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the "Why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*.

Guckelsberger, C. 2020. *Intrinsic Motivation in Computational Creativity Applied to Videogames*. Ph.D. Dissertation, Queen Mary University of London, United Kingdom.

Hulatt, O., ed. 2013. *Aesthetic and Artistic Autonomy*. Bloomsbury Publishing.

John-Steiner, V. 2000. *Creative collaboration*. OUP.

Jordanous, A. 2018. Creativity vs quality: why the distinction matters when evaluating computational creativity systems. In *Proceedings of the 5th Computational Creativity Symposium at the AISB Convention*.

Liu, Y. et al. 2023. Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models. *arXiv:2304.01852*.

Luck, M., and d'Inverno, M. 1995. A formal framework for agency and autonomy. In *Proceedings of the First International Conference on Multi-Agent Systems*.

Markus, H., and Hamedani, M. 2007. Sociocultural psychology: The dynamic interdependence among self-systems and social systems. *Handbook of cultural psychology*.

Monro, G. 2011. The image in the age of computer agency. Unpublished preprint at: *gordonmonro.com/notes/docs/Monro_G_Impact7_paper_draft2.pdf*.

Montuori, A., and Purser, R. E. 1995. Deconstructing the lone genius myth: Toward a contextual view of creativity. *Journal of Humanistic psychology* 35(3):69–112.

Nakajima, Y. 2023. Task-driven autonomous agent utilizing GPT-4, Pinecone, and LangChain for diverse applications. *Blogpost at: https://yoheinakajima.com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications/*.

Rhodes, M. 1961. An analysis of creativity. *The Phi Delta Kappan* 42(7):305–310.

Runco, M. A. 2004. Creativity. *Annual Review of Psychology* 55:657–687.

Weiner, R. 2000. *Creativity and beyond: Cultures, values, and change*. SUNY Press.

3. Evaluation

A computer model to evaluate the coherence of characters' behaviour based on their emotional relationships

Rafael Pérez y Pérez
Departamento de Tecnologías de la Información
Universidad Autónoma Metropolitana, Cuajimalpa, Ciudad de México
rperez@cua.uam.mx

Abstract

This paper describes a model that evaluates the emotional coherence (EC model) in a narrative. A story can be defined as a sequence of actions. Each time an action is executed, new emotional relationships or conflicts between characters are established, or existing ones are modified. This interaction between characters generates what I refer to as an Emotional Trace (ET), i.e., a record of the emotional relationships and conflicts between characters as the plot progresses. The model described in this paper offers a methodology based on the analysis of the ET to determine the emotional coherence of a story. A prototype of the EC model was implemented in MEXICA and used to evaluate three scenes. Through an Internet questionnaire, a group of subjects were asked to evaluate the coherence of the same three scenes. The results of the study reported in this paper suggest that the EC model represents a solid first step to provide MEXICA, and other automatic storytellers, with the ability to evaluate the emotional coherence of its characters.

Introduction

The automatic generation of narratives is a relevant area of research in computational creativity. Through the years researchers have employed diverse techniques to generate stories, e.g., problem solving, planning, rule-based systems, genetic algorithms, and recently deep neural networks (Sharples and Pérez y Pérez 2022). These approaches have been useful to understand the role of characters' goals, authors' goals, combinatorial approaches, language models, and so on, in automatic narrative generation. But a story is not only a sequence of words and phrases, or a set of goals to be achieved by a character. It also comprises emotions, conflicts, fears, hopes, traditions, ways of interpreting the world, ways of communicating, ways of challenging our most entrenched ideas, among others. Thus, research in the automatic generation of narratives requires to explore mechanisms to include, at least, some of these social, emotional and cognitive dimensions.

Our research group has focused for several years on the study of the use of emotional relationships and conflicts between characters as a means to develop stories. The use of emotions in plot generators is not new. For example, TALE-SPIN (Meehan 1976) and MINSTREL (Turner 1993) use variables that characterize the emotional states of characters, and those variables are used as part of the conditions necessary to activate some goals. DAYDREAMER (Mueller 1987) goes further, by employing these types of variables to control the flow of the program, that is, to activate and deactivate goals during the execution of the program (for details on how these systems work see Pérez y Pérez and Sharples, forthcoming). However, I do not know of any system that works with emotional relationships and conflict between characters as a mechanism to progress a story action by action.

For many years, one of the main challenges in this area of knowledge has been the creation of systems capable of producing coherent narratives. One popular solution is to establish mechanisms that allow the programmer to ensure that the output that the system produces is consistent. A typical example is the use of human-designed story-structures, which provide a logical and fluent way to progress a plot. In this case, the production of a story consists in instantiating such story-structures. More recent programs, such as those based on deep neural networks, use statistical relationships to generate texts. The first versions of this type of systems produced sequences of words that soon lost coherence. Their main problem was the "lack of memory" of what had happened earlier in the plot. Therefore, as the text progressed, the sentences lost connections with each other. The development of Transformers (Vaswani et al. 2017) significantly helped to reduce this problem. But it is still necessary to continue working on this limitation. In our case, our system, known as MEXICA (Pérez y Pérez and Sharples 2001; Pérez y Pérez 2015), employs emotional links and conflicts between characters to build a story context that allows progressing the plot, action by action, in a coherent way, avoiding the use of predefined story-structures. This approach offers an alternative research path to those mentioned earlier for the production of coherent texts.

The mechanisms involved in the way we humans write harmonious narratives are much more complex than those represented by our systems. The study of what I refer to as *Automatic Story Coherence* (ASCO), i.e., the study of how to represent in computational terms those methods and knowledge structures necessary for the production of texts that humans classify as coherent, is relevant for computational creativity. In this paper I present a model that aims to contribute to understand better how to develop coherent narratives that avoids using predefined story-structures.

How did the idea of building the EC model come about?

As Montford and Pérez y Pérez explains (forthcoming), the construction of our storyteller MEXICA followed four main steps:

1. Development of a cognitive account of creative writing.
2. Transformation of that cognitive account into a computer model.
3. A detailed study of how each of the elements in the computer model interact, and how they manipulate and transform information.
4. Evaluation of the outputs produced by the systems and analysis of the relation between the output's features and the elements and parameters of the model.

Step one consists of carrying out a study on different theories of the writing process. In his cognitive account of creative writing, Mike Sharples (1999) consolidates and expands the work of researchers in creativity and writing. These are the cores ideas of his cognitive account. As we write, we are constantly switching between two states known as engagement and reflection. During engagement, the writer automatically generates sequences of actions that progress the plot. The typical example is when we daydream. During reflection, the writer evaluates the material produced so far and, if necessary, he modifies it. For example, if an action generated during engagement is not fully justified in the plot, during reflection new events that give meaning to such an action are introduced. In this way, the writing process is a constant cycle between engagement and reflection.

Step two consists of figuring out how to represent in computer terms the ideas expressed by Sharples, and how to complement his cognitive account. The result of step two is the computer model of engagement and reflection (ER model) (Pérez y Pérez 1999). One of the most interesting contributions of the ER model to Sharples' account is the representation and use of emotional links and conflicts between characters as a way to progress a plot (Pérez y Pérez 2007).

Steps three and four are essential to evaluate the theoretical aspects of the ER model and its functionality. The evaluation of the narratives produced, as well as the study of the relationship between the characteristics of such narratives and the parameters of the system, are critical to continue developing this project.

While working in step four I detected the need to develop the model of emotional coherence (EC model). The following illustrates how during the developing of a plot MEXICA produces situations that require to assess the emotional coherence of characters. The system starts and during engagement generates the following sequence of actions: The princess cured the knight's wounds; the knight was grateful towards the princess; the princess and the knight fell in love. Then, the program switches to Reflection. MEXICA begins the analysis of the story in progress and realizes that the plot requires justifying why the princess heals the knight. Using its knowledge base, the program discovers that if in a previous action someone hates the knight and then hurts him, then the princess's action is justified. In this way, MEXICA inserts the action 'Someone hated the knight and then hurt him' at the beginning of the narrative. The next step is to instantiate the unknown character. The program has several routines to perform that task. The most important of them is inspired by the study of human improvisation, where it is stated that whenever possible a character should be reintroduced in a story in progress (Johnstone 1989). In this way, MEXICA reintroduces the princess (who is the only option), which leads to the following sequence: the princess hated the knight and then hurt him; the princess cured the knight's wounds; the knight was grateful towards the princess; the princess and the knight fall in love.

This description feels wrong. If the princess hates the knight, it does not make sense that then she heals him and falls in love with him. That is, based on common-sense knowledge, there is a contradiction between the behaviour of the princess and the emotions she feels towards the knight. I refer to this situation as *emotionally incoherent*. Thus, it is evident the necessity of designing mechanisms that allows the system to detect this type of sequences of actions. None of the theories about creativity and writing studied to build the ER model contemplated a similar circumstance. In this article I report the solution that I designed to extend the ER model and, in this way, provide MEXICA with the ability to evaluate the coherence of the emotional behaviour of the characters.

What is an Emotional Trace?

As explained earlier, MEXICA develops a story action by action. An action has associated a set of preconditions and consequences, in terms of emotional relations and conflicts between characters. Emotional relations can have a positive or negative valence. The precondition of the action Character-A cured Character-B is that B must be ill or wounded (a conflict); the consequence is that B is very grateful towards A (an emotional link with a positive valence). Each time an action is executed, new emotional relationships or conflicts between characters are established, or existing ones are modified. This interaction between characters generates what I refer to as an *Emotional Trace* (ET). Thus, an Emotional Trace is defined as the record of the emotional relationships and conflicts between characters as the plot progresses. This work claims that the ET is important to evaluate the coherence of a narrative.

Let me elaborate this idea. Imagine a story with three characters, Carmen, Julia and Maria. As the story progress, the first interaction between characters establishes the initial state of their emotional trace. If the consequence of the first action is that Carmen is fond of Julia (an emotional link with a positive valence), one expects that in the following events that positive relation between them continues or strengths. In this way, as the story progress, it makes sense that Carmen helps Julia to solve a problem, or that they become best friends. I refer to this situation as a *positive emotional trace* between Carmen and Julia. Similarly, if the consequence of the first interaction between characters triggers an aversion from Carmen to Maria (an emotional link with a negative valence), one expects that in the following events that negative relation between them continues or strengths. In this way, it makes sense that Carmen and Maria end being enemies, or that they sabotage each other's goals. I refer to this situation as a negative emotional trace between Carmen and Maria.

Thus, given a partial story where two characters have developed a positive emotional trace, the coherence of a narrative is disrupted when an action with negative consequences between these two characters is added to the tale. For instance, if Carmen and Julia are best friends, it does not make sense that out of the blue Carmen betrays Julia. Similarly, a partial story where two characters have developed a negative emotional trace, the coherence of a narrative is disrupted when an action with positive consequences between these two characters is added to the tale. For instance, if Carmen and Maria are enemies, it does not make sense that in the next action Carmen asks Maria to be the godmother of her son. I refer to this type of situations as *disruption of the emotional trace*. Thus, in this work, a narrative is emotionally coherent when there are not disruptions in any of its ETs.

However, good stories are full of descriptions where a character betrays his brother, or falls in love with an enemy, and so on. These scenarios are not accidents; the change in the emotional trace between characters have a narrative purpose. To keep the coherence of the tale, this change in the emotional relations between characters must be justified. For instance, imagine a story where Carmen and Paul are rivals, i.e., they have a negative ET. But one day Carmen realises that Paul risked his life to safe her young nephews in danger. As a result, now Carmen is fond of Paul. In this case, the action where Paul saves the children has the purpose of changing the emotional relation from Carmen to Paul. Now, it makes sense that the plot continues with a scene where Carmen is friendly towards Paul. I refer to this type of actions as *transitional actions*, because they help to make a coherent transition from a negative emotional trace towards a positive one, or vice versa.

The Emotional Coherence Model (EC model)

In this model, all characters who participate in an action have associated an attribute that can be set to one of two values: Proactive (P) or Reactive (R). A proactive character represents an actor that performs an action with the aim of provoking an emotional reaction or conflict in himself or in the other character. By contrast, a reactive character represents an actor that reacts to the action executed by a proactive character or, in some occasions, that reacts to the designs of fate (for example, in the event of an accident). Given an action that involves two characters, e.g., Character-A ACTION Character-B, there are four possible configurations of proactive and reactive characters:

- Character-A is proactive and character-B is reactive, represented as (ApBr).
- Character-A is reactive and Character-B is proactive, represented as (ArBp).
- Character-A is proactive and Character-B is proactive, represented as (ApBp).
- Character-A is reactive and Character-B is reactive, represented as (ArBr).

In the following, I provide details about how the EC model works. For the sake of clarity, I assume that an action only includes two characters, A and B. So, the consequences of an action can trigger emotional links and/or conflicts, either in one of the characters, or in both characters.

Case when the consequences of an action trigger emotional links and conflicts in one of the characters. For this analysis I assume that A is proactive and B is reactive (ApBr). In this case, there are two possible scenarios:

(1) When the post conditions of an action only trigger emotional links and conflicts from the proactive Character A towards the reactive Character B. That is, the consequences of the action performed by the proactive character only produces an emotional reaction in itself; character B does not react towards A. This instance is known as consequences on Character A (CA). An example is "A was jealous of B", whose consequences triggers a negative emotional link from A towards B, represented as -CA. In the same way, the consequence of the action "A admired B" is that character A establishes a positive emotional link towards B, represented as +CA.

(2) When the post conditions of an action only trigger emotional links and conflicts from the reactive Character B towards the proactive Character A. That is, the consequences of the action performed by the proactive

character only produces a reaction in the reactive actor, so A does not react towards B. This instance is known as consequences on Character B (CB). An example is "A insulted B" whose consequence triggers a negative emotional link from B towards A, represented as -CB. A second example is "A cured B", whose consequence triggers a positive emotional link from B towards A, represented as +CB.

Given these two scenarios, the EC model includes three rules for a coherent emotional trace.

Rule 1. Given an Emotional Trace (ET) and an action to be used to continue the story (ACT_{t+1}), when the consequences of the action ACT_{t+1} are negative (-CA or -CB), then the ET cannot include positive emotional relations from the Proactive Character A towards the reactive Character B (see figure 1). Otherwise, the story is classified as emotionally incoherent.

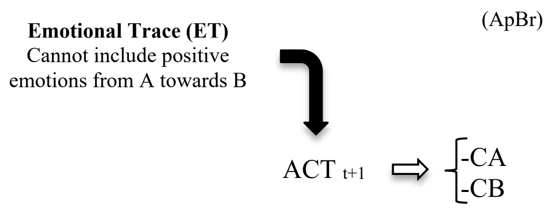


Figure 1. Representation of Rule 1.

Let me elaborate this idea. To perform the action A wounded B, that triggers a negative emotional link from B towards A (-CB), makes sense if A and B are rivals, i.e., if the ET between A and B includes a negative emotional relation from A (proactive) towards B (reactive). By contrast, the same action A wounded B does not make sense if A and B are best friends, i.e., if the ET between A and B includes a positive emotional relation from A (proactive) towards B (reactive).

Rule 2. Given an Emotional Trace (ET) and an action to be used to continue the story (ACT_{t+1}), when the consequences of the action ACT_{t+1} are positive (+CA or +CB), then the ET cannot include negative emotional relations from the Proactive Character A towards the reactive Character B (see figure 2). Otherwise, the story is classified as emotionally incoherent.

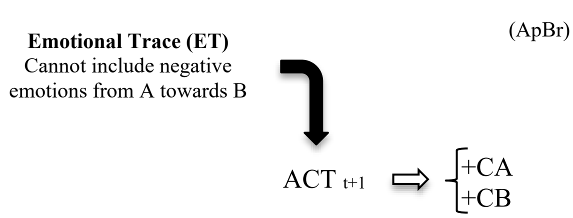


Figure 2. Representation of Rule 2.

Thus, the action A buys a birthday present for B, whose consequence triggers a positive link from B towards A (+CB) makes a lot of sense if they are best friends (i.e., if their ET includes positive emotional links between the characters); on the other hand, if character A and character B are rivals (i.e., if their ET includes negative links), this action is illogical.

Rule 3. Given an Emotional Trace (ET) that includes both, a positive and negative relations from Character A towards Character B (this situation is known as clashing emotions), the consequences of the action ACT_{t+1} can be either positive (+CA or +CB) or negative (-CA or -CB). If the consequences are positive, the negative emotional relations from the Proactive Character A towards the reactive Character B in the ET are not considered anymore (it is as if the system eliminated them). If the consequences are negative, the positive emotional relations from the Proactive Character A towards the reactive Character B in the ET are not considered anymore.

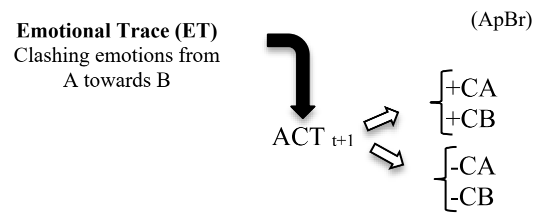


Figure 3. Representation of Rule 3.

So, if character A develops clashing emotions towards character B, then the behaviour of character A can be driven either by their positive emotions towards B, or by its negative emotions. But when a decision is made, it cannot be changed. During the rest of the tale character A must behave in the same way.

These three rules also work when A is the reactive character and B is the proactive character (ArBp). The difference is that the emotional relations that need to be checked in the ET are those from B (proactive) towards A (reactive).

Case when the consequences of an action trigger emotional links and/or conflicts in both characters. Some actions might produce a reaction in both characters, i.e. as a consequence of performing the action, both characters develop emotional relations and/or conflicts towards each other. This case is known as consequences on A and B (CAB). CABs can have four combinations of proactive and reactive characters: ArBr, ApBr, ArBp, ApBp, and the consequences might be positive, negative, or a mixture of them (see table 1). Let me examine each situation.

(i) The consequences of an action are always logical when both characters are reactive (ArBr). In other words, in this case there cannot be an emotional incongruity. An example is "A and B had an accident", where nobody has the intention to harm the other.

(ii) In a situation where A is proactive and B is reactive (ApBr), the action's consequences are initially evaluated as a CA and then as a CB. That is, first the coherence of all emotions from A towards B are evaluated, considering that A is the proactive actor. Then, the coherence of all emotions from B towards A are evaluated, considering that A is the proactive actor. Here there is an example. The consequences of the action "While healing his wounds, Character A falls in love with Character B" are that Character A develops a strong love emotion towards B (+CA) while Character B reacts with gratitude towards A (+CB). Clearly, B has no intention whatsoever in this action. So, this action only makes sense if character A has positive emotional links towards B, no matter what B's emotional links towards A are.

(iii) The situation where ArBp is a mirror case of the previous one. So, the proactive and reactive characters are inverted, but the process is the same.

(iv) For the situation where both characters are Proactive (ApBp), the process works as follows. The coherence of all emotions from A towards B, and those from B towards A, are evaluated, considering that A is the proactive actor. Next, the coherence of all emotions from A towards B, and those from B towards A, are evaluated, considering that B is the proactive actor. In other words, (ApBp) can be picture as first processing the consequences as (ApBr) and then processing them as (ArBp). For instance, "A and B became best friends". In this example, both characters have the purpose of strength their relationship. The action only makes sense if both characters A and B do not have any negative emotional relations or conflicts between them. Similarly, "A and B insulted each other" only makes sense if characters A and B had negative emotional links or conflicts between them.

Proactive and reactive characters	Description
ArBr	The consequences of an action are always logical.
ApBr	Considering that A is the proactive actor, first the coherence of all emotions from A towards B are evaluated; then, the coherence of all emotions from B towards A are evaluated.
ArBp	This is a mirror case of the ApBr.
ApBp	The coherence of all emotions from A towards B, and those from B towards A, are evaluated, considering that A is the proactive actor. Next, the coherence of all emotions from A towards B, and those from B towards A, are evaluated, considering that B is the proactive actor.

Table 1. The four cases of proactive and reactive characters when the action has consequences on A and B (CAB).

Testing the EC model

A prototype of the EC model was developed to evaluate how these ideas work. Given an initial sequence of actions, we asked the program to verify if different endings have or not an emotional coherence. The initial sequence introduces a king, a princess and a knight (for the sake of clarity, in these examples I employ friendly texts rather that the raw inputs that the system uses):

The king was the proud father of the princess. For many years, the king and the knight had hated each other. However, the knight and the young princess fell in love…

This description establishes the following emotional links: a strong positive emotional relation between the king and his daughter; a strong hated-base negative relation between the king and the knight; and a strong positive love relation between the princess and the knight. Next, the system evaluates three scenes that share the same initial sequence but had different endings. For the sake of clarity, the endings are marked in bold.

Scene 1:
The king was the proud father of the princess. For many years, the king and the knight had hated each other. However, the knight and the young princess fell in love. ***The king killed the princess; then, he killed the jaguar knight.***

This scene ends with the king killing the lovers. So, the king is the proactive actor and the lovers are the reactive actors. After analysing this story, the system generates the following report:

```
Result of the analysis of illogical actions.
The story includes the following 1 illogical action(s):
    Action 7 -> King killed Princess
Explanation:
  7 King killed Princess -> Earlier KING had established a positive relationship with PRINCESS.
```

The system reports that the king and the princess have a positive relationship, so her murder does not make sense (the system only reports uncoherent actions). To keep this story coherent, the solution is to include a transitional action, e.g., The king felt betrayed by the princess. The king can only feel betrayed if he and the princess have a strong positive emotional relation. The consequence of this action is that the king develops negative emotions towards the princess. That will justify the murder.

In the following example, the king feels betrayed by the princess (this is a transitional action) and then he kills the knight (the king is the proactive character and the knight is the reactive character). But this time, the princess (proactive character) wounds the king (reactive character), and then she cures him:

Scene 2:
The king was the proud father of the princess. For many years, the king and the knight had hated each other. However, the knight and the young princess fell in love. ***The king felt betrayed by his daughter. So, he killed the knight. The princess wounded the king. Later, the princess cured the king.***

Result of the analysis of illogical actions.
The story includes the following 1 illogical action(s):
Action 10 -> Princess cured King
Explanation:
10 Princess cured King -> The consequences of this action are not in concordance with the emotional trace from PRINCESS towards KING. Earlier PRINCESS had established a negative relation with KING.

At the beginning of the story the king and the princess had a positive relationship. The murder of her lover produces that the princess establishes a negative relation with the king. So, she has clashing emotions towards her father. Then, the princess decides to wound her father. So, the behaviour of the princess establishes that the negative relation in their ET is dominant over the positive one. Now, she needs to be congruent with that behaviour. But she is not (she cures her father). That is why the system reports that curing the king does not make sense.
In the last example, the king feels betrayed by the princess (this is a transitional action) and then he attacks the knight (the king is the proactive character and the knight is the reactive character). As a reaction the knight (proactive character) wounds the king (reactive character). Then, the princess (proactive character) kills the knight (reactive character) and heals the king (reactive character):

Scene 3:
The king was the proud father of the princess. For many years, the king and the knight had hated each other. However, the knight and the young princess fell in love. ***The king felt betrayed by his daughter. So, he attacked the knight. The knight wounded the king. The princess killed the knight and then she cured the king.***

In this case, the princess has strong positive and negative emotional relations towards the king (because he attacked her lover) and also towards the knight (because he wounded her father). The princess reacts negatively towards the knight and positively towards her father. Both reactions follow the rule 3 of coherence, so the system does not report any problem.

What do people think about these endings?

Through an Internet questionnaire, a group of subjects were asked to evaluate the coherence of the same three scenes introduced in the previous section. The aim was to study if the criteria followed by the EC model resembles the criteria employed by a group of human judges when performing the task of analysing the consistency of a sequence of actions.
The questionnaire was in Spanish and was divided into four sections. The first part explained the objective of the study and requested the age, gender, and the last academic degree of the participant. In the three remaining sections, each of the scenes from the previous section were presented, and the participants were asked two questions to evaluate the coherence of the behaviour of the characters that participated in the closing of the scene. The possible answers to each question were a numerical value between 1 and 5, where 1 represented "very little coherent" and 5 represented "very coherent". In addition, for each question, participants were asked to explain why they granted that grade.
The questions for scene 1 were:
- Does it seem coherent to you that the king kills the princess?
- Does it seem coherent to you that the king then kills the knight?
The questions for scene 2 were:
- Does it seem coherent to you that the king kills the knight?
- Does it seem coherent to you that the princess first wounds the king and then heals him?
The questions for scene 3 were:
- Does it seem coherent to you that the king attacks the knight, and in response to the attack the knight injures the king?
- Does it seem coherent to you that the princess then kills the knight and cures the king?

39 subjects answered the questionnaire. 53.7% of them identified themselves as males, 43.7% as females and 2.6% as nonbinary. The range of ages covered from 19 to 76 years. 51.3% had technical degree or were undergraduate students, 20.5% had a bachelor degree, and 28.2% had a postgraduate degree. In the following analysis I include descriptions made by the participants about why they decided to grant a specific coherence value. Those descriptions were translated from Spanish to English by this author with the help of an automatic translator. For reasons of space, I only include a small sample of these comments.

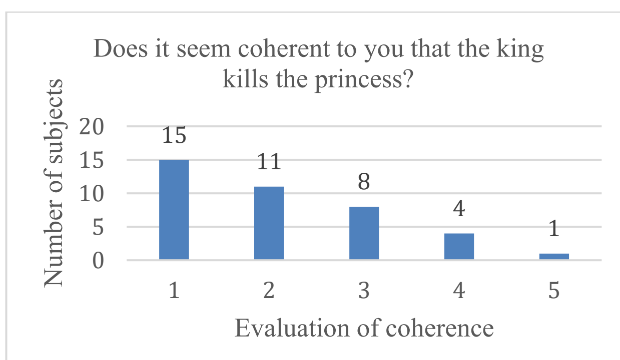


Figure 4. Answers to the first question of scene 1.

Results of scene 1. Figure 4 shows the results to the first question of scene 1. 66.7% of the subjects ranked the action where the king killed the princess with a value of 1 or 2, i.e., most people thought that it did not make sense. 20.5% of the subjects ranked the action with a value of 3, i.e., they were unsure about the coherence of the action. 12.9% of the subjects ranked the action with a value of 4 or 5, i.e., they classified this action as coherent.
Some participants that ranked the action as incoherent (i.e., they assigned values of 1 or 2) explained their reasons to give such values as follows:
- “Because the king loved the princess, she was his daughter”
- “Because if you love your daughter you will never hurt her”
- “If he is proud of her, it means that there is also an affective bond that would make it impossible for him to hurt her, why would he kill her then?”
- “If the father was proud of her daughter it doesn't make sense for him to suddenly kill her.”
- “It seems to me that the king's love for his daughter, the princess, surpasses his hatred for the knight, and if the princess is happy with the knight, no matter how much the king hates the knight, it does not seem to me that there is any justification for killing the princess.”
These comments suggests that the positive emotional relationship described in the text between the king and the princess is the main reason why the action where the king kills the princess seems incoherent. This view matches rule 1 of the EC model that establishes that if the king (the proactive character) has positive emotional relations towards the princess (the reactive character) and then he performs an action with negative consequences for her (e.g., killing her), such an action is classified as incoherent. Thus, the evaluation generated by our computer model coincides with the opinion of the majority of the subjects.

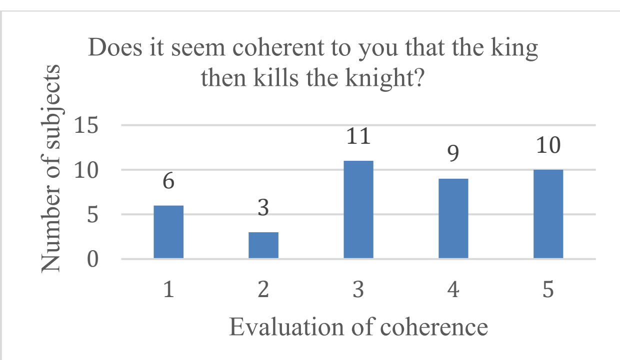


Figure 5. Answers to the second question of scene 1.

Figure 5 shows the results to the second question of scene 1. 48.7% of the subjects ranked the action where the king killed the knight with a value of 4 or 5. That is, most people thought that it did make sense. 28.2% of the subjects ranked the action with a value of 3, i.e., they were unsure about the coherence of the action. 23.1% of the subjects ranked the action with a value of 1 or 2, i.e., they classified this action as uncoherent.
Some participants that ranked the action as coherent (i.e., they assigned values of 4 or 5) explained their reasons to give such values as follows:
- “Yes, it is consistent because it was mentioned at the beginning of the story that the king and the knight hated each other, then, due to that hatred, it can be deduced that he killed him.”
- “the king and the knight hated each other”
- “because he hates him”
These comments seem to be consistent with rule 1 of the EC model. In this case, the king (the proactive character) has negative emotional relations towards the knight (the reactive character) and then the king performs an action with negative consequences for the knight. As a result, the action is classified as coherent. Thus, the evaluation generated by our computer model coincides with the opinion of the majority of the subjects.

Results of scene 2. Figure 6 shows the results to the first question of scene 2. 58.9% of the subjects ranked the action where the king killed the knight with a value of 4 or 5, i.e., most people thought that it did make sense. 28.2% of the subjects ranked the action with a value of 3, i.e., they were unsure about the coherence of the action. 12.8% of the subjects ranked the action with a value of 1, i.e., they classified this action as uncoherent. None ranked the action with a value of 2.
Some participants that ranked the action as coherent (i.e., they assigned values of 4 or 5) explained their reasons to give such values as follows:
- “He hates the knight and he killed him because he felt his daughter betrayed him. It makes sense.”
- “Because of his hatred towards him”
- “Because he hated it. He could have killed him out of jealousy or out of hate itself.”
- “For many years the king and the knight hated each other.”
- “Because of the unbridled hatred that the King felt for the knight.”

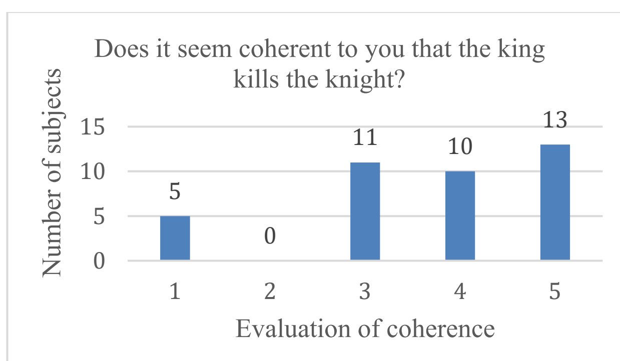


Figure 6. Answers to the first question of scene 2.

Like in the previous question, these comments seem to be consistent with rule 1 of the EC model. In this case, the king (the proactive character) has negative emotional relations towards the knight (the reactive character) and then the king performs an action with negative consequences for the knight. As a result, the action is classified as coherent. Thus, the evaluation generated by our computer model coincides with the opinion of the majority of the subjects.

Figure 7 shows the results to the second question of scene 2. 48.7% of the subjects ranked the behaviour of the princess with a value of 3, i.e., most people were uncertain about the coherence of her actions. 33.3% of the subjects ranked the situation with a value of 4 or 5, i.e., they thought that the actions made sense. 17.9% of the subjects ranked the actions with a value of 1 or 2, i.e., they classified the princess' behaviour as incoherent.

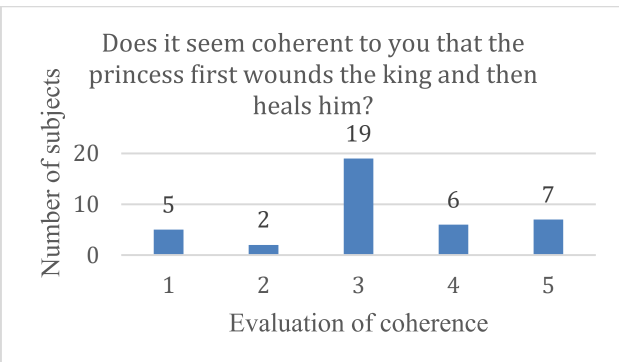


Figure 7. Answers to the second question of scene 2.

Some participants that were uncertain about the coherence of the action (i.e., they assigned a value of 3) explained their reasons to give such a value as follows:

- "The intermediate events that justify the change of posture of the princess are missing. One can imagine them, but the story does not deliver them..."
- "It could be because he was her father and she could have hurt him by chance."
- "It seems possible to me, since I imagine a scenario in which the princess reacts impulsively but then her love for her father and her possible guilt make her cure him. However, it seems unlikely to me."
- "Both things can happen without contradicting each other, if the affective bond between the king and the princess is strong, any situation can be overcome. However, that would detract from the love story between the knight and the princess."

-"Yes and no, this would depend on how the character of the princess has been presented to us throughout the tale, since that would be very important to define what she would have done, since after injuring him she could flee or remain in pain by the death of her knight but repentant for hurting his father"

Scene 2 shows a scenario where the princess develops conflictive feelings towards her father; she hates him and she loves him. The clashing emotions appears to cause confusion to the participants. In their comments, they acknowledge that the princess has reasons to act in either way. Rule 3 of the EC model states that if the princess develops clashing emotions towards the king, the behaviour of the princess can be driven either by their positive emotions towards the king, or by its negative emotions. But when a decision is made, it cannot be changed i.e., during the rest of the tale the princess must behave in the same negative way towards the king. Situations where the princess regrets from her previous actions, as suggested by some of the subjects, are not contemplated in the current version of the model. These results suggest that the basis of rule 3 are correct but this rule might require to be updated (see the discussion section).

Results of scene 3. Figure 8 shows the results to the first question of scene 3. 83.9% of the subjects ranked the interaction between the king and the knight with a value of 4 or 5, i.e., most people thought that it did make sense. 12.8% of the subjects ranked the action with a value of 3, i.e., they were unsure about the coherence of the situation. 7.7% of the subjects ranked the action with a value of 1 or 2, i.e., they classified this action as uncoherent.

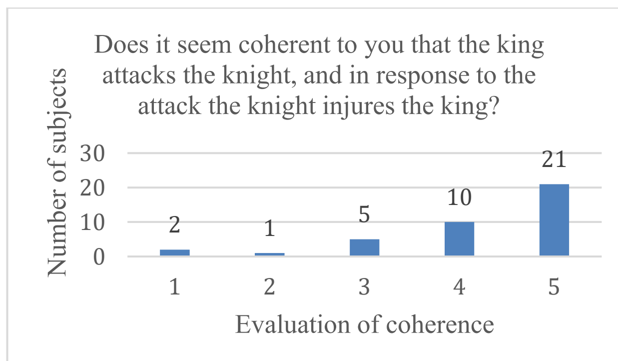


Figure 8. Answers to the first question of scene 3.

Some participants that ranked the action as coherent (i.e., they assigned values of 4 or 5) explained their reasons to give such values as follows:

- "This sequence is very logical and very coherent. If the two hated each other, it means the king's response to the knight and the consequence of his attack (he was wounded by the knight)."
- "they were enemies"
- "Yes, he could hurt him as a defence to the attack"
- "They are rivals, the fact of being attacked by your rival gives you an excuse to hit back."
- "The king was carried away by emotions and attacked, the knight defended himself."

Again, these comments seem to be in accordance with rule 1 of the EC model. Thus, the evaluation generated by our computer model coincides with the opinion of the majority of the subjects.

Figure 9 shows the results to the second question of scene 3. 48.8% of the subjects ranked the princess' behaviour with a value of 1 or 2, i.e., most people thought that it did not make sense. 28.2% of the subjects ranked this situation with a value of 3, i.e., they were unsure about the coherence of the actions. 23% of the subjects ranked the action with a value of 4 or 5, i.e., they classified these actions as coherent.

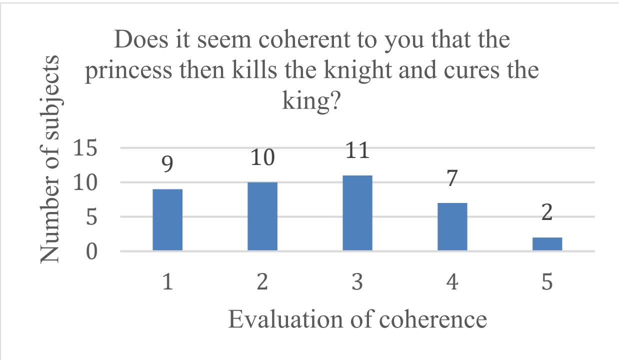


Figure 9. Answers to the second question of scene 3.

Some participants that ranked the action as incoherent (i.e., they assigned values of 1 or 2) explained their reasons to give such values as follows:
- "If she is in love with the knight she would not kill him. It makes sense to me that she heals the king"
- "Because she loved the knight"
- "No, the knight defended himself from an attack; if the princess really loved him she wouldn't have killed him. At the most she would have turned him into a toad."
- "Well, the love she felt for the knight would not cause her to have killed her lover"
- "Normally the princess, being supposedly 'in love' with the knight, would have made him escape, while she would stay to heal the king"

These comments suggest that the fact that the princess reacts negatively towards the knight is not necessarily a problem; rather, the strong consequences of her reaction (killing her lover) is what makes the princess's actions being perceived as illogical. The fact that the princess is in love seems to constrain in the eyes of the evaluators what constitutes an adequate negative reaction. None of the comments argues against the princess healing the king.
In scene 3, the princess experiences two different clashing emotions. The princess was in love with the knight and, at the same time, she hated the knight because he injured her father. Simultaneously, the princess loved her father, but also hated him because he attacked her lover. Thus, the princess had strong positive and negative emotional relations towards the king and also towards the knight. Rule 3 of the EC model states that if the princess develops clashing emotions towards the king and towards the knight, the behaviour of the princess can be driven either by her positive or negative emotions towards each of those characters. But when a decision is made, it cannot be changed. Scene 3 satisfies this rule; the princess decides to act nasty towards the knight and compassionate towards the king. So, the model classifies this scene as coherent. What the model does not evaluate is whether the princess's reaction is excessive, which seems to be the main reason why subjects evaluated the princess' actions as uncoherent.

Discussion

This paper describes a model of emotional coherence for narrative generation that has been instantiated as a computer program. One of the main challenges in research on automatic narrative generation is how to produce coherent texts. There are several characteristics to consider when one evaluates the consistency of a tale. For example, the adequate structure of the narrative, the congruence between the behaviour of the characters and their goals, their personalities, their roles in the story, the social norms represented in the story, among many others. This work focuses on a model capable of evaluating the coherence of the characters' behaviour based on their emotional relationships. This type of analysis is necessary because the development of a story necessarily implies the construction of emotional relationships between its characters that change over time. Such changes must be consistent, otherwise the story loses cohesion.
The results of the study reported in this paper suggest that the EC model has a solid foundation. Most of the explanations given by the study's participants about why they evaluated characters' actions in a certain way seem to coincide with the fundamentals that drive the model. Comments made by the participants in the clashing emotions conditions suggest the need to consider the following situations originally not contemplated:
(i) In scenes like the one where the princess hurts the king and then heals him, it is necessary to relax the rigidity of rule 3 in order to allow including situations such as when the princess regrets her previous actions and therefore it makes sense that later she heals her father.
(ii) In scenes like the one where the princess decides to kill the knight and then cure her father, it is necessary to consider whether the consequences of her actions towards the knight are proportional to the emotional relationship between them.
(iii) The model should also evaluate if the emotional links between characters that are important to understand characters' behaviours are clearly shown in the story. This will help to prevent readers' confusions.

The EC model works in contexts where emotional relationships between characters, such as friendship or rivalry, are explicitly represented in the system and shape the behaviour of the characters. That is, negative emotions between characters foster actions with negative consequences, while positive emotions between characters foster actions with positive consequences. Typically, this behaviour is stable and only changes through transitional actions. For other narrative contexts, for example, stories about a serial killer where the assassin has no emotions towards his victims, or at least not the type of emotions that most of us would feel in a similar circumstance, the EC model requires to be expanded.

The work described in this paper has been incorporated into MEXICA, our storyteller, where it is used during engagement, where the story is progressed, during reflection, where the story is analysed and modified, and during the evaluation phase, where the system evaluates its own output. The system also includes a module where the user can type and evaluate any story, as long as the text follows the rigid format the MEXICA employs. The EC model also has other potential applications. For example, it can be used as part of the fitness function in systems based on genetic algorithms. It can also be used in videogames or interactive storytelling.

Narratology distinguishes between the story level (or content) and the narrative discourse (or expression). The model of emotional coherence works at the story level. However, once the system has evaluated the consistency of the story, it can use resources such as ellipsis, at the narrative discourse level, to produce more interesting narratives.

The results of this work suggests that MEXICA is useful to test, modify and expand the original concepts and theories employed as framework to build the computer model of engagement and reflection. There are many interesting aspects related to story generation which have not been represented computationally. The vast limitations of all existing systems make evident the need to explore new mechanisms. It is clear the necessity of finding new ways to produce narratives automatically, that go beyond the way we commonly use techniques such as problem solving, genetic algorithms, and recently deep neural networks. The incorporation of perspectives, methodologies and knowledge arising from the humanities and the social sciences will undoubtedly revitalize the area. Automatic narrative generation, and computational creativity in general, will greatly benefit from it.

References

Johnstone, K. 1989. *IMPRO Improvisation and the Theatre*. Methuen Drama.

Meehan, J. R. (1976). The Metanovel: Writing Stories by Computer. PhD Dissertation, Yale University.

Montfort, N. and Pérez y Pérez, R. (forthcoming). Computational Models for Understanding Narrative. Revista de Comunicação e Linguagens. https://rcl.fcsh.unl.pt/index.php/rcl

Mueller, E. T. (1987). Daydreaming and computation: A computer model of everyday creativity, learning, and emotions in the human stream of thought (Technical Report UCLA-AI-87-8). Doctoral dissertation, Computer Science Department, University of California, Los Angeles, CA.

Pérez y Pérez, R. (1999). MEXICA: a computer model of creativity in writing. Doctoral dissertation, University of Sussex, UK.

Pérez y Pérez, R. (2007). Employing Emotions to Drive Plot Generation in a Computer-Based Storyteller. *Cognitive Systems Research*. Vol. 8, number 2, pp. 89-109.

Pérez y Pérez, R. (2015). A Computer-based Model for Collaborative Narrative Generation. *Cognitive Systems Research*, 36-37, p. 30-48.

Pérez y Pérez, R. and Sharples, M. (2001) MEXICA: a computer model of a cognitive account of creative writing. *Journal of Experimental and Theoretical Artificial Intelligence*. Volume 13, number 2, pp. 119-139.

Pérez y Pérez, R. and Sharples, M. (forthcoming). *An Introduction to Narrative Generators: How Computers Create Works of Fiction*. Oxford: Oxford University Press.

Sharples, M. (1999). *How we write? Writing as creative design*. London: Routledge.

Sharples, M. and Pérez y Pérez, R. (2022). *Story Machines. How Computers Have Become Creative Writers*. London: Routledge.

Turner S. R. (1993). *MINSTREL: A computer model of creativity and storytelling*, PhD Dissertation, University of California LA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, December 2017, pp. 6000-6010.

Reviewing, Creativity, and Algorithmic Information Theory

Dan Brown[1] and Max Peeperkorn[2]
[1] Cheriton School of Computer Science, University of Waterloo, Canada
[2] School of Computing, University of Kent, Canterbury, UK
dan.brown@uwaterloo.ca, m.peeperkorn@kent.ac.uk

Abstract

We connect critical review and analysis of creative objects to a recent domain-independent creativity assessment framework by Mondol and Brown (2021a; 2021b). Reviewing is interesting for at least three reasons. Reviews are time- and space-limited, unlike other tasks. Reviews are a creative task about creative tasks, and that meta-creativity is interesting to consider theoretically. And reviews cause communication and learning; the various actors (the primary creator, the reviewer, and the reader) interact in complex ways. We show how Mondol and Brown's framework connects to the process of review, and show how topics like summarization, contextualization and learning fit within an algorithmic information theory frame. We also give some interesting examples, such as analysis of conceptual art and concert reviews, as computation tasks. We finish by showing that (as is often true of algorithmic information theory ideas) it is hard to fulfill our objectives with practical systems, due to uncomputability or intractibility issues.

Introduction

Reviewing, either of a creative product like a concert or novel, or of an academic product like this paper, includes several goals that the reviewer must satisfy. First, the goal is to provide an overall assessment of the quality of the product in the first place: is the paper good, should other people come to see the further offerings of the performance, and so on. Second, the reviewer summarizes the product: what is the relationship between the main characters of the novel, say. Third, a good reviewer offers contextual information, for example: how does the object fit within or expand its genre, or was the performance better than expected? Academic reviews also come with creative improvements and recommendations from the reviewer.

Reviewing is a computational task. With this lens, we see inputs, background knowledge, execution time constraints, memory use constraints, and output size constraints, and can identify what a critic does, what makes a good critic, why different critics will identify different aspects of a performance, how readers can interpret a review, and more.

Review is also creative. Readers may be delighted by a review's cleverness, or by its surprising insights. A reviewer may unexpectedly connect properties of the manuscript with the author's biography, or may examine an oeuvre in a novel way, or identify how the work expands knowledge.

We approach review via a computational creativity (CC) framework and focus on review quality and creativity. Additionally, we consider certain aspects of how reviews are created (in particular, their limited creation times), how reviews affect readers, and how reviews can form a communication channel between a creator and a potential audience. In emphasizing these perspectives on reviews, we reconstruct the 4P's framework (Rhodes 1961), adapted to CC by Jordanous (2016). That is, reviews are themselves a Product, created by a Producer, via a Process, for an audience (Press).

We focus our theory on the algorithmic information theory approach to computational creativity evaluation of Mondol and Brown (2021a; 2021b). That framework offers a genre-independent definition for key concepts like novelty, quality, typicality and more. However, our focus on review also extends Mondol and Brown's work, in that we focus on properties of the reviewer: what program is the reviewer executing? What background information does the reviewer have? How much time does the reviewer have in which to make a judgment? Does the reviewer have prior knowledge or expectation about who will be reading the review?

We also focus on the reader of the review: their background and ability to interpret the review. As such, our approach focuses on the review as a complicated communication intermediary between a creator and a consumer. For agents to talk about creative objects, we need to formalize how they convey their opinions. Thus, key to our understanding of review is the relationship between a reviewer and the reader. An expert reviewer may identify aspects of a creative object that a naive reader cannot understand; meanwhile, a naive reviewer may not be of use to an expert reader, who already understands the work and wants to assess if, for example, a concert of a beloved piece is worth attending. One example of a reviewer might in fact be the creator of an object, summarizing it as a teaser for potential readers; this still fits within our frame, especially if the creator is knowledgeable about potential readers.

Our goal here is to document how the Mondol and Brown framework can be applied to the overall task of critical review. We discuss what it means to be a "good" review or reviewer. And, we identify some serious challenges relating to the Mondol and Brown framework in the first place, given that their frame expects that the quality of an object is connected to its mathematical *logical depth* and *sophisti-*

ISBN: 978-989-54160-5-9

cation, two hard-to-estimate quantities, and that the formulation by Mondol and Brown does not restrict the work of a reviewer in approximating the information found in an object. We connect our work to computational creativity by discussing previous efforts to shape critical review as a CC task, and to describe how individual reviewers might bring their own perspectives to the task of review. Our paper contributes to overall understanding of how artistic objects form a means of communication between producers and their audience, and the role intermediaries play in that communication act.

Background

Here, we give two different pieces of background to our project. First, we describe how review has previously been portrayed in the computational creativity literature. Then we give the necessary mathematical background for this overall work, focusing on the Mondol and Brown framework. We finish this section by giving the notational framework in which we embed the review task in terms of its various actors.

Review and CC

Few papers discuss *review* in computational creativity, but two ICCC papers (Fisher and Shin 2019; Roberts and Fisher 2020b) do. Both discuss an early effort by Stiny and Gips (1978) connected to algorithmic aesthetics, where they connect design directly with criticism, and so the full Stiny and Gips model consists of two parts, a structure for design algorithms, and a structure for criticism algorithms. In this paper, with the Stiny and Gips model, we specifically refer to its criticism component. Stiny and Gips base their model on Craik's general model of thought (Craik 1943), which consists of a *receptor* that produces a description of what it senses in the world, a *processor* that transforms the description for the *effector* to produce an observable response.

In the Stiny and Gips model (based on Roberts and Fisher (2020b)), receptor function R takes object α and contextual information I_α to produce description δ. Processor P, in the Stiny and Gips model, is an algorithm informed by an aesthetic system, which is a set of algorithms that form the aesthetic criteria, and a memory of contextual information I_m. P takes description δ and contextual information I_m, and outputs, in addition to the original description δ, the best interpretation ι and numerical aesthetic evaluation ϵ. The effector E is a function that takes δ, ι, and ϵ to generate review χ.

Fisher and Shin (2019) identify review as a separate creative task and highlight that critics are part of larger creative ecosystems. They argue for the importance of computational critics and identify five desiderata for a critic. The computational critic should 1) understand the medium, 2) emphasize the authorial intent of an artifact, 3) reason about the creator's output and its relationship to the subject, 4) situate the artifact in social and historical contexts, and finally, 5) gauge the response of the readers and viewers to the critique. Besides the standard CC criteria for creativity (Boden 1992; Ritchie 2007), they analyze essential dimensions involved with critique: authority, authenticity, explainability, and interpretability. Additionally, because of the role critics play in society, they address the ethical concerns for when computational critics are implemented. The follow-up paper (Roberts and Fisher 2020b) further explores the Stiny and Gips model and adjusts their approach by formalizing their desiderata. In particular, they introduce a justification γ, as another output of the analysis algorithm P and an additional input to E. We extend their approach by focusing specifically on the computation that happens in both the reviewer and in the reader, and looking at properties of all of these agents.

Scientific review has been identified as a domain for implementing computational critics because it provides excellent grounding in a specific context (Fisher and Shin 2019). A first attempt at a computational scientific critic is *pReview* (Roberts and Fisher 2020a). A recent paper (Yuan, Liu, and Neubig 2022) discusses automating scientific review as a Large Language Model (LLM) task; that paper is more oriented around the practical LLM techniques to do this automation, and identifies some key desiderata.

Algorithmic Information Theory and Creativity

Our frame for analyzing creativity is the algorithmic information theory (AIT) framework of Mondol and Brown (2021a; 2021b). They give a Product-focused definition of basic concepts in creativity, including value, typicality, and novelty. In their framework, all of these concepts are based on properties of a Turing machine program whose output is the digital objects under study: an object s is *valuable*, for example, if there are short programs whose output is s, but where all of them require long runtimes to execute. Here, we give a brief introduction to the Mondol and Brown framework; interested readers are referred to the full paper (Mondol and Brown 2021b) for more detail. The standard textbook on Kolmogorov complexity (Li and Vitányi 2019) gives more complete definitions than those that follow.

Kolmogorov complexity We study digital objects, represented unambiguously. Given a universal Turing machine U that can generate any computable object s, the *Kolmogorov complexity* $K_U(s)$ is the length of the shortest input P^* for which $U(P^*) = s$. We ignore details of U, and often describe $K(s)$ without reference to U, speaking of the execution of P, not U. The runtime of program P is the number of Turing machine steps before P halts (and is infinite if it does not halt). The conditional Kolmogorov complexity $K(s|y)$, is the length of the shortest Turing machine which, on input y, outputs s: this quantity measures how similar s and y are, or how much knowing y allows us to compress the string s.

$K(s)$ alone is insufficient to identify if s is creative. Random sequences have Kolmogorov complexity very close to their length with high probability. The n-bit string 0^n has very low Kolmogorov complexity, at most $\log_2 n$. Both are not of creative value. Instead, Mondol and Brown use two other concepts in AIT as evidence of an object's creative value: *logical depth* and *sophistication*.

Logical depth The *logical depth* of s is the minimum runtime of programs with output s and with their length close to $K(s)$. Specifically, $ld_c(s) = \min_{P:U(P)=s,|P|\le K(s)+c} \mathrm{runtime}(P)$, for some small parameter c. Objects with short, fast-running programs are not deep (they are highly and trivially compressible and decompressible). Objects with only long programs are not deep (they are random). Mondol and Brown show that high-quality objects can be compressed, but their decompression is slow: non-random parts of the object must be painstakingly reconstructed. Consider, for example, a painting where the positions of key objects are described by a short-to-describe algorithm that requires a long time to execute: there is structure in the painting, but it is hard to tease out. The logical depth model of value says that for an object to be of high quality, there must be substantial and complex work embedded in the object. By contrast, Schmidhuber (2010) has argued that beauty (which he treats as similar to quality) is connected primarily to being of short description, regardless of the required runtime of a generation algorithm.

Sophistication The *sophistication* of s comes from a two-part representation of digital objects. An object is defined by giving the class of objects for which it is typical (to its *model*, M), and the information required to describe the specific element in that class. (There is often deliberate ambiguity between M, the program for the model and $L(M)$, the class generated by M.) Valid models are typically restricted; one straightforward requirement is that models are Turing machines that halt on all inputs, or Turing machines that can generate any output. With such a restriction on valid models, then, the sophistication of a string is $soph_c(s) = \min_{(M,d):|M|+|d|\le K(s)+c, U(M,d)=s} |M|$. It is the shortest model for which the two-part representation comes close to optimally encoding the important details of s. The model encodes the category of objects for which s is a *typical* example; typicality is of course a standard desideratum for computational creativity (Ritchie 2007).

Sophistication is not an easy concept. The restriction to models that are total functions removes the universal Turing machine U as a valid model, since (without it) $soph_{|U|}(s) \le |U|$ regardless of what s is, since the universal Turing machine, run on the shortest program for S, will yield s. Instead, the model framework requires the identification of a computable class of objects with a relatively short description that includes s as a *typical* member, and then the details that identify s from all of the other class members.

Both of our previous examples of non-valuable objects have short models. A random string s with high $K(s)$ has as its model a constant-length "print" program p that just outputs its input, so s then has a two-part code of length $|s|+|p|$. By contrast, we need to be a bit more careful in describing a repeated pattern. Consider the string $s = 0^k$, where the number k has $K(k) \approx \log_2 k$ (that is, the binary string k is uncompressable). It is easily modelled by a constant-length program p that on input of a binary number x, outputs 0^x; then the two-part code (k, p) has total length very close to $K(s)$, so p is a good model for s. Neither of these strings, hence, is sophisticated: they both have short models.

By contrast, let s have fairly high Kolmogorov complexity, yet compressible by complex programs. It cannot be simply represented by its Turing machine representation as input to U, but we can build a model of similar size to $|p| = K(s)$ whose language only contains s: this machine ignores its input, and implements a universal Turing machine running p, which always halts and outputs s. The existence of this machine shows that the sophistication (assuming this machine is a valid model) cannot be substantially larger than $K(s)$, but it could be smaller if the two-part representation allows the model M to encode all of the compressible information in s, while allowing the data d to be uncompressible.

Key to the relationship between sophistication and value is the important part of the two-part representation, M, which represents the regularizable information found in all outputs of the model. By contrast, d, the data, is random information ultimately not relevant to the meaning of the model. A naive consumer of s may not be aware of the inherent information of s, and assumes s is less sophisticated than it is (mistaking s for more random data), or they may not understand s sufficiently to compress it fully. This challenge is the connection between review and the Mondol/Brown framework.

Compression, lossy compression, and generation The Kolmogorov complexity of an object s defines how much information is in s, by giving an optimal compression for s. If s is logically deep, no short program speedily generates s: the only speedy generators of s require longer descriptions. If we restrict to fast generators, we can only use models that do not fully understand the information found in s.

If we restrict to programs that are both fast and short, we cannot generate all of a logically deep string s. Instead, we can only lossily represent s. Let $s'_{t,n}$ be the closest approximation to s that we can obtain by fast, short programs: that is, $s' = \mathrm{argmin}_{s'} K(s|s')$, where s' searches over all machines of size at most n bits and with runtime at most t steps. If n is close to $K(s)$, and the runtime is kept smaller than the logical depth, then s' may be able to represent some surface features of s, but cannot identify the valuable pieces of s. However, if the runtime is kept moderate, and the programs must be short, we can still potentially explore some small piece of the logically deep core of s, which may offer some hint that the whole object is valuable.

Generation is also key to review. If we assert s is a "typical" example of a genre, that implies that a generator G for that genre, run on random inputs, would yield an object of similar quality and appearance. (We note that we will often use the more informal term "genre" to refer to the class of objects created by a generator, instead of the more common algorithmic information theory term "class"; in part, this is because we want to focus on creative objects.) For example, if G generates typical romance stories, then the parametrization might indicate the names of the characters or their occupations, as well as some arbitrary details about the story, then G could generate that new story. We only claim to *understand* s when we can describe such a generator; further, if on other random inputs, G's output does not fit the genre, then G is a bad representation of the class. Consider a general-purpose compression system, like the Lempel-Ziv

(Ziv and Lempel 1978) algorithm. It may compress s, but if run on a different input, it is likely to generate a completely different type of output than s. As such, it is not a good model for s.

Compression and generation are very difficult tasks to understand for simple objects because good generators of small objects must be much larger than the objects themselves; this yields a situation in which to analyze the quality of an object, we must instead consider a collection of objects of a type or make dramatic restrictions on what is a valid model; see Brown and Mondol (2021) for more details. Table 1 gives a summary of AIT concepts used in this paper.

Review and algorithms

In our formulation, A creates the object s, B uses s as one of its inputs and creates a review r of s, and C uses r to contribute to its understanding of s, and whether or not to further investigate s. For example, A might be a movie studio, creating a new movie; B writes a newspaper review of the review, and C decides whether the movie is worth going to on the basis of the review, in addition to learning from the review.

Different reviewers may discover novel aspects of s; for example, B_1 may focus on the brushstrokes of a painting, while B_2 focuses on the biographical details of the creator and B_3 focuses on just giving service journalism about the exhibition housing it. All of these might be more or less useful to readers; we will discuss this issue later in the paper.

Review as a multi-part CC task

A review of a creative object includes several parts (Fisher and Shin 2019). There is a summary of the object, situating it in the domain from which it comes: for a novel, perhaps talking about the characters and their relationships. The review will include the reviewer's assessment of *quality*, either in textual form, or as a numerical rating. It can include framing information: how the genre has changed, or what the reviewer adores or despises. The review may describe how the new object alters one's understanding of the field, or include biographical information about either the creator or reviewer. They can also suggest improvements; this is appropriate for academic papers, but could apply to any object: a recipe tester might identify missing flavours, or a musical reviewer might identify awkward lyrics or harmonies.

Review as a computational and creative task

Each reviewing subtask is computational: the input to an algorithm is the object s, and the reviewer's process in moving from their own knowledge and mental state to the textual review is the execution of an algorithm. Further, review can also be seen as *lossy compression*: in summarizing a piece of music, one gives enough description that a reader has some better idea of the piece of music than they had before reading the review, and (if it is well-prepared) a better estimate of the quality of the piece of music than they had before reading the review. Formally, consider an object s under review, created by a creator A. Assume that A is a Turing machine computing a total function (A halts on all inputs), and $s = A(d)$, for some input d that comes from the outside environment of A. Program A is not known to any other observer. Reviewer B is also a Turing machine that is given s as input, and computes a review $r = B(s)$. B may have some framing information as part of s (the name of the composer of a song, the date of a performance, the title of a poem, or an artist's statement); in fact, it can be valuable to consider $s = (x, y)$, where x is the main object and y is its framing information (Charnley, Pease, and Colton 2012). Review r should include (possibly not well separated) each of the parts we have described: the summary, estimate of quality, and contextual framing for the object s. Then, the review r is sent to the reader C, potentially along with the same framing information available to B, and it is used by C to help update its understanding of the object s, without C being able to see the object s or its creation by A directly. See Figure 1 for a conceptual model of this process.

Conceptual model of review in AIT

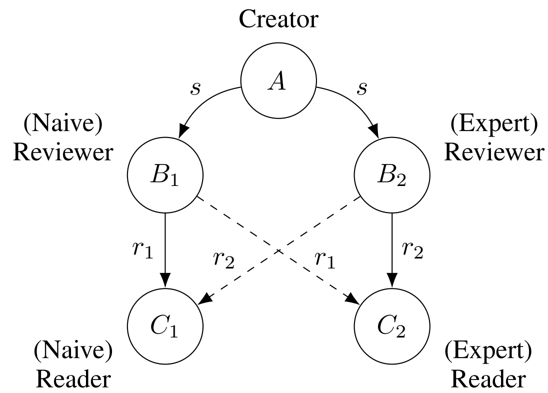


Figure 1: Here, s is the object under study, r_1 is a general review, and r_2 is an expert review. The usefulness of a review r to a reader C is dependent on its prior model of s.

Review is creative: in addition to identifying the properties of s, an expert reviewer can tantalize and educate their readers. Famous movie reviewers like Pauline Kael and Roger Ebert were considered masters of the craft, due to their encyclopedic knowledge of the movie industry and its customs, and other critics like Susan Sontag can join a diverse range of fields in giving context to s and why it matters. Hence, reviews can be assessed for typicality (is this recognizably a review?), novelty (does it identify unexpected things about s?) and value (does it show important aspects of s, or focus only on surface features of s?), and similarly, B can also be assessed (though the distinction between B and r can be hard to identify).

How review differs from other computational tasks

One other key condition governs review, and makes it simpler than other forms of critique (such as, for example, large literary biographies): reviews are created in a short time frame and must fit in a short space. One might be tasked with an 800-word review of a 2-hour concert, due three hours after the concert ends, or one might be handed an 8-page paper

AIT idea	notation	quick summary	meaning for reviews
K-complexity	$K(s)$	shortest program whose output is s	not useful
conditional K-complexity	$K(x\|y)$	how much information is shared between x and y	measuring novelty and typicality
logical depth	$ld_c(s)$	runtime of short programs for s	one measure of value
sophistication	$soph_c(s)$	length of shortest model for s	another measure of value
model	M	program M that outputs s on a specific random input d	way of describing s
good model	M	program M whose length is close to the sophistication of s	appropriate way of describing s

Table 1: AIT concepts used in this paper

and a 2-month deadline in which to review it. B is thus limited to Turing machines that always halt: they must halt in a restricted time t and with their output r restricted to a limit λ. This limitation is interesting: if $t < |s|$, the reviewer cannot examine all of s, and if $\lambda < K(s)$, then it is impossible to fully describe the object (in terms of describing a program that generates s in its entirety). Further, if $t < ld(s)$, then it is not possible to verify the correctness of a short program for s (let alone find that program in the first place!). If $\lambda < soph(s)$, then we cannot give a good model for s.

Another difference, which we explore later in the paper, is that review is connected to the reader of the review's pre-existing understanding of the object under study. Review is a step in *learning*: if a new reader can be brought up to speed about the genre by reading good straightforward reviews of masterpieces of the domain that include descriptions of important milestones in the genre, this is a valuable service to that reader, as it allows them to build better models of both s and its genre. If the reader is already knowledgeable, then the needed review to help them better understand a complicated new piece of work is more complex.

Review and AIT

Review fits messily with AIT. A review with a short execution time cannot identify that an object is sophisticated, for two reasons. First, even if B knew the program and input that A used to generate s, it cannot run that program in a small amount of time. If s is sophisticated, then the program/data pair (p, d) that documents its sophistication is both long and slow to execute. Even if (p, d) generates s, a simpler program might also have generated it: the creator of s may have toiled to create s, but not made a valuable object.

Instead, all review tasks can only be approximated. The reviewer B can consider the object s within their expertise, and can find evidence for novelty and value, but cannot be guaranteed of success. As B specializes, they may be more prepared to find such evidence, but at the loss of broad applicability, particularly given that B has a time limit.

Aesthetic evaluation and AIT

To identify value, we estimate logical depth or sophistication. Novelty has a full description in the Mondol/Brown framework as well (Mondol and Brown 2021b), based on how much a new object differs from a corpus of members of the same genre. A novel object is both familiar (since it can be placed in an existing model), but also unfamiliar (it is distinguishable from other objects generated by the model in not just random, meaningless ways). This framework lets us not view a sonnet as "novel" when placed in a set of objects that are all paintings, for example. Quality (as novelty, typicality, and value) is estimated by describing what an object is in the context of a good model for that object and its class, and why it is exemplary of a sophisticated model.

Summarization and AIT

To summarize effectively, one should identify the model from which an object is a typical example, and indicate the ways in which the model does or does not fully satisfy the non-random information found in the object. This relates to quality estimation: the critic identifies the model from which an object comes, and also the random parameters for that model. For example, "It's a Jackson Pollock painting, with the paint spills in these positions, with these colors." Again, novelty estimation is separate from the summarization process, and thus the task of review is not just summarization.

Contextualization and AIT

Contextualization is included in an object's model: if we know about a good model M for s, it cannot create surprise, since s is then a typical example of M's outputs. This requires that the model does not accidentally move outside the genre on random inputs. We discuss this topic in the examples later, but a straightforward example is Duchamp's "Fountain" (which was a urinal): obviously, a very short model can ignore its input and generate a sufficient description of the piece, but it will not generate any other Duchamp-readymades, nor distinguish them from fakes. By contrast, a much larger model that presents Duchamp's Dadaist background and the scenarios in which he worked might generate other examples of Dadaist conceptual art. Obviously, a short review cannot give a full description of such a model, but can present information necessary to describe how it differs from models, like "print" programs.

The role of the reader of a review

Our discussion so far has treated review as a very quickly created description of an object, focused on quality estimation and model identification. But reviews communicate: the readers of the review have their own tastes, their own goals, their own knowledge, and needs to extract from the review useful information about the object s.

This complex relationship requires unpacking. These three actors are not playing a complex game of Telephone: instead, a smart reviewer learns about both creator and reader to ensure that useful information gets passed between these two actors. B does not not merely build a short description of the model A used to create s, but also needs to pick the best way to describe this model, given what C already knows about objects like s. If C is an expert on the topic of objects like s, then the updates B needs to give to the internal model C has may be of a very different class than the model description given by B to a naive reader. See Figure 1 for a sketch of this process.

This observation highlights a key problem with the Mondol/Brown framework: it equates quality with logical depth or sophistication, removing the chance for a reader to have preferences and tastes. Their "complexity above all" frame does not allow for this person to like romance novels and that person to like conceptual art. Instead, for any object, the goal is to group it with objects of similar kind, and establish how well the new object extends that group: how much s expands the set of valid members and how much computation is built into its non-random parts. We must expand our understanding of the role of a reviewer and consider both learning and the Press component of the 4Ps of understanding creativity (Jordanous 2016).

Learning

A clear goal of a review is to allow C to better understand s and estimate its value. Our definition of value is about required computational effort and expanding our understanding the class of s, so, the B wants C to change its model of the class of objects s comes from. That is to say, C reads the review r, and as a result, its model for s, which C has not seen, changes, as does its assessment of the value of both s and its class. This is what is meant by learning: we modify our knowledge of how objects and concepts relate. If r does not allow C to either assess s or to alter its understanding of the class, then C did not learn anything useful, and the review was useless *to that reader*, though it might be useful to others. If B is writing without an audience in mind, then their optimal choice is to express observations that will be clear to any reader, implicitly assuming that C has no prior background in the subject under consideration, and that without encountering s or r, C will know nothing about s. If C is already an expert on objects like s, then B's job in writing a review is to focus C's interest on aspects of s that highlight quality or novelty, and on subtle reasons why s differs from previous members of its type.

Estimating the quality of a review Consider review r of object s. A natural way of assessing the quality of the review r is to ask how much r simplifies s: that is, what is $K(s) - K(s|r)$? But if the information of r consists overwhelmingly of random details about s, then having the review will not inform C about anything useful to assess s. Instead, the review must help build a more accurate model of the generator of s for r to have actually been valuable to C, or it must represent the logically deep components of s.

Review r can be even worse than just offering random information about the object under study: it can make false claims! Consider a review of a concert that misidentifies the set list, or a review of an art exhibition that gives false contextual information about the painter. The review provides no information to further the reader's understanding, and the model the reader C brings to the show may be less accurate (finding s more atypical) than before. We will consider this a bit further below when we look at the reader's experience.

To estimate the quality of review: If M is a good model of s, then by conditioning on r, we can estimate how many additional regularities are captured after observing r, which is $K(M) - K(M|r)$. Conversely, the irregularities of s are modelled with $K(s|M)$. We could fit the model M with r to estimate if the model improves as it captures more irregularities, that is $K(s|M) - K(s|M, r)$. We can separate information about s in r into two parts: those that connect to M and those that connect to its random input.

Alternatively, we can use the logical-depth frame to also explore the quality of a review: let $ld_c(s|r)$ be the minimum runtime of a program for s whose length is at most $K(s|r)+c$. This computes how much runtime is needed to give a short description of s given r. If $ld_c(s|r) \ll ld_c(s)$, then the review r captures critical information found in s, and is therefore a good review. If instead the required runtime has not changed much, then while r may capture information about s, it does not conveyed much of value.

Estimating the quality of a review to a reader Again, we must also consider the quality of a review to a specific reader. If C does not read Czech, then a brilliant Czech review will do nothing to help the reader. Moreover, there is insufficient information in r to teach the reader to read Czech. We must consider how reading r affects C, and see how information is transferred.

Let $K(M|C)$ be how much information C needs in order to create a good model M, without having seen the review r. If $K(M|C)$ is high, then s is complex and the reader is unprepared; if $K(M|C)$ is small, the reader is prepared. In both cases, r may help C change its model. Suppose that M_1 is a model that does a good job of explaining previous examples that C has seen, and that M is a model that explains both those previous objects and also s. $K(M|M_1)$ is the added novelty brought to bear by the creation of s. The key quantity under consideration, then, is $K(M|M_1) - K(M|M_1, r)$. That is, how much information is created in C by reading r that is relevant to M? A detailed review that establishes a small component of M, and how it changes as a result of s, may be of help to an advanced reader, but may contain much less overall information than a general review. However, that general review may offer little actual new content to a reader C.

We may impose a time limit on C's execution as well. It is probably inappropriate to allow C enough time in understanding r to learn a new language. If we put a time limit t' on C's computation using r as an input, then let $M' = C(r)$ be the new model that C has after running for t' steps on input r; if $K(M|M', C) < K(M|C)$, then C has learned something useful about s from the review r in its short time.

C can also have knowledge about specific reviewers, and can learn to trust a particular reviewer B to give good advice. Our model does not really handle this circumstance, which has been brought to our attention by a peer reviewer (whom we hope enjoyed our work); however, C can identify which reviewers' overall quality estimates most track with its own choices. It is worth recalling that in the Mondol/Brown frame, since quality is an absolute quantity, there is less accounting for individual taste; we discuss this below when we consider limitations of our framework.

Communication through a naive reviewer A review can *still* be useful to an expert reader even if the author of the review is not an expert. B may explicitly notice features of s that are novel even to an experienced reader. If an artist has changed their colour palette, B may not know that the shift has happened, while highlighting A's colour choices. C can then update their model of A's work, even though B does not explicitly represent the change. A related example of this phenomenon might also be if someone noted the D-S-C-H motif in "Rejoice in the Lamb" (Britten 1943), which is an homage to Dmitri Shostakovich, without explaining the four-note sequence. Communication happens between A and C even though B is not aware of the content.

Creative reviews, creative reviewers

Reviews themselves are creative objects. A review r of a creative object s has *value* by describing s (giving information about the model that generates s, or about the logical depth and novelty of s). But it can have novelty and value in that process. Consider reviewers B_1 and B_2; if both identify elements of s showing it is of high quality, but those identified by B_1 are more often known by people in the audience than those identified by B_2, then to a typical reader, B_2's review will be novel, and as such have much more value than B_1's. A clever reviewer discovers new things to enjoy about a piece of creative work, and then shares that joy. Reviews can also provide delight to their readers in and of themselves; not only can they highlight the creativity and model-breaking natures of a new object, but they can just be objects of creative value in their own right. It is a challenge to separate these aspects of a review's quality from the overall analysis of creativity.

Creative reviews can also suggest improvements. AIT does not offer an easy way to give small corrections; if these corrections take up the bulk of a review, they will not change the underlying model very much (since one can use an existing model augmented with "at line X, change word Y to word Z" commands). Reviews that describe the underlying model for s that the reviewer B believes A has used could be made richer, and improve the overall value or novelty s.

We can also explore the creativity of the reviewer, not just the review; while the Producer perspective is not an obvious use case for AIT (which might be expected to focus on products, since it connects to properties of objects), one can either analyze the program that B executes in its review process, or one can focus on a collection of reviews by B, to see whether a single creative review is an accidental flash of genius or represents consistent excellence in a reviewer's work.

Examples and limitations of the approach

Here we explore some real-world examples of review and how they fit within our conceptual framework.

Conceptual music and art

Consider the iconic piece 4'33" by John Cage, in which the performers sit for four minutes and thirty-three seconds making no deliberate sounds. A performance of this work is hard to describe in a single object s, but the piece can be "summarized" easily. But if we look at it with an eye towards AIT, and in particular, towards sophistication, it is insufficient to model it as a "print" program whose input is "make no deliberate sounds for 273 seconds." This model ignores the awkwardness of sitting in a room with other humans where normally one expects to see music performed in the normal way. Instead, to properly summarize, one needs a model that, on random inputs, yields typical performance experiences for even this ostensibly simple piece, differing only in random details. Such a model is likely impossible, even for 4'33", to present in a short review. B must include descriptions, likely to improve C's understanding, enough to push its model closer to the truth of what the piece is. To describe the novelty and value of the experience, C's experience with conceptual music would need to have been pre-estimated by B (for a naive reader, this piece utterly alters their experience of what a concert is; a seasoned reader would understand what differentiates performances).

Similarly, looking at Duchamp's "Fountain" mentioned earlier. The short "it's an early twentieth-century urinal in an art gallery" review does not offer enough information on why this piece provokes such ire among gallery-goers, and certainly does not allow the reader to assess whether it improves or worsens the show. By contrast, a review that describes the state of early 1900s sculpture, and describes how "Fountain" expands the art gallery experience, is potentially a much better review, giving a reader a better sense of what other provocative conceptual art would be like. The "it's a urinal in a gallery" reader will be more prepared for other examples of other nouns to replace "urinal", but will not have a reason to understand scatological conceptual art in general; while a reader given a provocative review that describes "Fountain" might find, for example, Andres Serrano's 1987 work "Piss Christ" (a photo of a crucifix in a vial of the artist's urine) less surprising.

A conceptual artist's process can also be the focus of the works; consider Roman Opalka's paintings of the "numbers from 1 to infinity", where the artist's project was to paint consecutive numbers to represent the passage of time. Here, one might review either the individual paintings, or the process itself, in either case, one would again contextualize the creator's practice within the genre of conceptual art.

Finally, Sol Lewitt's work, which consists of short algorithmic descriptions of exactly how someone is to create the object. It is possible that a model that creates algorithmic art might be created by a reader, after reading a review. The review might describe several specific choices made by the person implementing the algorithm, to give a sense of what a different typical implementation looks like, and the complexity needed to properly understand Lewitt's work.

Searching for a relevant explanation for conceptual art, then, is a computational task of finding evidence for quality, valuable summarization, and contextual information. Even for "simple" conceptual pieces, there may be much to discover. Again, review is not merely lossy encoding of s: it is lossy encoding of s and how to understand it.

Concert reviews, short and long

A concert review offers an opportunity for both conveying important details about the performance (date, time, venue, set list) and also individual aspects about the performer and their genre. In a short review, the reviewer may still convey core performance details to naive readers, but these details are obvious to a well-prepared fan. A review that goes into detail about what was amazing might help out a novice reader to understand what demonstrates successful performances. While also focusing on specific details that made the concert different, allowing the expert to see why there is novelty and not just value in the performance.

A much simpler review (or preview) can also still yield important information about genre and quality. One of us once saw the Hilliard Ensemble, a Renaissance vocal quartet, describe their next piece as: "Late Tallis, early Byrd". These four words prepared the listener for the upcoming piece, while also making it clear that the singers are English (the piece they were about to present was French). Like so many other cases of short sentences, these four words provide much context and much model shaping to a prepared listener, and no context at all to a novice (they might even confuse one into believing the piece was English!).

Limitations of our approach

Fundamentally, our approach *describes* what reviewers need to do in the process of assessing quality, typicality, and novelty of a creative object, but the actual project of creating, within a time limit, a high-quality short review is complex.

Reviewing good objects is hard Reviewers assessing a sophisticated object face a challenging task: they must identify, in a short time, why the object comes from a large, slow model M. If B is itself highly sophisticated, it may be able to zero in on certain subfeatures of the object under study, and identify why these are consistent with the object overall being of good quality: for example, they could describe a single object in an exhibit, giving the reader enough information to better understand what to look for on a time-unlimited tour through an installation. In this way, the reviewer may spend serious effort to identify bits of information from the creator program A, but nonetheless, the review is productive, since $K(A|r) < K(A)$.

It is also possible that a reviewer unprepared to analyze the high-quality object cannot, in the short time allowed, explain the features of the object. In this case, while B fills r with true information about s, it will not be very useful to C. In the AIT sense, this means that the reviewer might describe some random bits d that are the input to the creator program A, not pieces of the structure of A itself. While these contribute to $K(s)$, they are not central to understanding s, and the reader of the object is not better prepared to encounter other objects of the type.

Since reviews are time limited, no critic is a general-purpose critic. Instead, a sophisticated, specialized critic, when handed an object it is well posed to review, can identify high-quality parts of the object efficiently, and is preset to describe the model in language that a reader will understand. A claimed general-purpose reviewer must spend some of its time chasing down blind alleys. We could theoretically handle this situation by creating teams of reviewers (the equivalent in computational terms of parallel programming), but the most sensible thing is to assign reviewers by expertise and by awareness of the identities of the readers.

Reviewing bad objects is hard Reviewing unsophisticated objects is also hard. If an object comes from a simple, fast model, it can still be extremely hard to verify in a short amount of time. A high-quality reviewer might have learned features of high-quality examples of a genre without sufficient awareness that these features are not the important items to find; if every good painting by a particular artist uses a lot of blue paint, it is easy to highlight this surface feature and incorrectly assigning high-quality to trivial new objects. Mondol and Brown (2021b) also discuss the *charlatan* phenomenon, when clever agents knowingly assign high-quality estimates to poor objects by describing complex programs with long runtimes to compute s; this can be seen as (fake) evidence of logical depth. Naive readers can easily be confused by such reviews into overestimating the object, while experienced readers still must identify errors in a review that claims a junky object is a work of genius.

An incorrect review r of a bad piece of art has almost no effect on a reviewer's actual understanding: the information gain of $K(M|C) - K(M|C, r)$ will be modest, since r gives minimal information towards M, a good model for s. Still, since a bad piece of creative work has low $K(M)$ to begin with, an incorrect review may accidentally give information toward an initial simple model; this is part of why we focus on the absolute number of bits in $K(M|C) - K(M|C, r)$.

AIT shows the challenge in review We end this discussion of AIT and reviews by noting that our message is not hopeful. Reviewers who can detect complexity quickly are rare, and we cannot verify high logical depth and sophistication in short runtimes since a short, fast program may also exist where a short, slow program has already been found. It *is* sometimes possible to properly explain why an object is trivial, but a critic can be caught up in enthusiasm for the trivial work of a beloved creator or a charlatan, and give an explanation that (falsely) highlights perceived complexity.

This difficulty comes down to the twin dilemmas of the Mondol/Brown aesthetic theory: objects are valuable if they embody much work, and such objects appear more trivial (random) to naive consumers until they are explained with appropriate models. (By contrast, the Schmidhuber (2010) approach focuses solely on K-complexity; it, too, is hard to tease out since simple strings may still appear complex unless one knows the short algorithm that explains them.) In a time-limited review, it is challenging or impossible to explore this work, and to properly explain these models. Instead, special-purpose reviewers and experts can only approximate this aesthetic lens.

Future work

We have described how reviewing fits with algorithmic information theory. The challenge is to make our insights practical: almost all areas AIT touches find either huge runtimes or uncomputable results. For example, identifying good models requires that not much smaller models work well, and that the model input is truly uncompressible; computing logical depth requires knowing $K(s)$ and the runtime of machines computing s. This is beyond a general-purpose algorithm with reasonable runtime for any interesting input, meaning that AIT largely provides an abstract restatement of a number of normal computational or human tasks. There are also a few specific concerns that connect to this specific application of AIT: notably, real-world algorithms for building seemingly creative objects seem huge, and that any form of embodiment can cause us to question what “the object under study” is, and whether it has a unique identity. Further work will address the queries below about practicality, parameterization, and embodiment.

Connections to machine learning

Large language models (LLMs) and other large machine learning models show some of the complexity of review, as they require a truly enormous number of parameters before giving reasonable language creation. If such a a system is the “model” in the sense of our paper, then $K(M)$ will always be huge: it is the size of the code for the LLM; the data part of the two-part code is the input (prompt) to the model. This, though, requires extremely long strings s for $K(M) < K(s)$; such strings cannot be well explored by a time-limited reviewer. We still require much assessment to figure out for what kinds of input the AIT frame can work.

Embodiment

Another challenge with linking AIT and review is the phenomenon of sensory embodiment (Guckelsberger et al. 2021). Do B and C interact with the same object s? We have assumed that s is properly represented by a single digital string, but B and C may perceive it differently. B and C may either or both experience disability and inexactness in their ability to perceive s. If B watches a concert from the front row, the experience described in r may be inaccessible to C from a back row. If B is colour-blind, r cannot help C learn about the subtle choices in shading that A made.

Another way in which embodiment (and other concerns about physical and memory limitations) affects AIT and review is the finiteness and sequencing of memory. In AIT, $K(s|y,x) \approx K(s|x,y)$; it does not matter much in which order the two objects x and y appear. But encountering two different large and complex objects will have different effects if they cannot both be stored; the more recently-experienced object may have more details in memory, while the more distantly-experienced object may have had a more fundamental effect on the internal model the reader has of the category from which s, y and x all come. The key effect here, then, is *forgetting*, which we propose to discuss in future work.

Conclusion

We present review of creative objects as a computational creativity task, using Mondol and Brown’s framework as a starting point. In our discussion, review is recast as quickly identifying features of a good model explaining an object. For a simple object, the features identified in a review can either be among the few interesting (non-random) aspects of the object, or might be simply surface features differentiating the object from similar ones, but which are also random. For complex objects, the best features to identify in a review tease out the complexity at the heart of the objects; unfortunately, these features can be extremely hard for a reviewer to identify in a short space and quick review period.

We have briefly discussed the ways in which targetting reviews for their audiences has an AIT formulation, and have also described why assessing creative and uncreative objects are both hard tasks. While our approach does not yield practical implementations, it gives a proper theoretical underpinning for a central task of the creative world.

Acknowledgments

The work of D.B. is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. The work of M.P. is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds. We are grateful to Tiasa Mondol for useful conversations, and to a somewhat anonymous reviewer, whose peer review comments were extremely helpful in improving this manuscript.

References

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Britten, B. 1943. *Rejoice in the Lamb. Op. 30*. Boosey and Hawkes.

Brown, D. G., and Mondol, T. 2021. On the problem of small objects. *Entropy* 23(11):1524.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In Maher, M.; Hammond, K.; Pease, A.; Pérez y Pérez, R.; Ventura, D.; and Wiggins, G., eds., *Proceedings of the Third International Conference on Computational Creativity*, 77–81. Third International Conference on Computational Creativity, ICCC 2012 ; Conference date: 30-05-2012 Through 01-06-2012.

Craik, K. 1943. *The Nature of Explanation*. Cambridge: Cambridge University Press.

Fisher, D. H., and Shin, H. 2019. Critique as creativity: Towards developing computational commentators on creative works. In *ICCC*, 172–179.

Guckelsberger, C.; Kantosalo, A.; Negrete-Yankelevich, S.; and Takala, T. 2021. Embodiment and computational creativity. In de Silva Garza, A. G.; Veale, T.; Aguilar, W.;

and y Pérez, R. P., eds., *Proceedings of the Twelfth International Conference on Computational Creativity, México City, México (Virtual), September 14-18, 2021*, 192–201. Association for Computational Creativity (ACC).

Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

Li, M., and Vitányi, P. M. 2019. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 4th edition.

Mondol, T., and Brown, D. 2021a. Incorporating algorithmic information theory into fundamental concepts of computational creativity. In *Proceedings of the International Conference on Computational Creativity*, 173–181.

Mondol, T., and Brown, D. G. 2021b. Computational creativity and aesthetics with algorithmic information theory. *Entropy* 23(12):1654.

Rhodes, M. 1961. An analysis of creativity. *The Phi Delta Kappan* 42(7):305–310.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Roberts, J., and Fisher, D. 2020a. preview: The artificially intelligent conference reviewer. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 665–668.

Roberts, J., and Fisher, D. H. 2020b. Extending the philosophy of computational criticism. In *ICCC*, 358–361.

Schmidhuber, J. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2(3):230–247.

Stiny, G., and Gips, J. 1978. *Algorithmic aesthetics: computer models for criticism and design in the arts*. Berkley, California: University of California Press.

Yuan, W.; Liu, P.; and Neubig, G. 2022. Can we automate scientific reviewing? *J. Artif. Intell. Res.* 75:171–212.

Ziv, J., and Lempel, A. 1978. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory* 24(5):530–536.

On Characterizations of Large Language Models and Creativity Evaluation

Max Peeperkorn[1], Daniel G. Brown[2], and Anna Jordanous[1]
[1] School of Computing, University of Kent, United Kingdom
[2] Cheriton School of Computer Science, University of Waterloo, Canada
m.peeperkorn@kent.ac.uk, dan.brown@uwaterloo.ca, a.k.jordanous@kent.ac.uk

Abstract

Incredible as they may be, Large Language Models (LLMs) have their limitations. While they generate high-quality texts, excel at stylistic reproduction, and tap into an immense pool of information, they can produce wildly inaccurate responses. The hype around LLMs led to them being characterized as "reasoning", "sentient", or "knowing" like humans. We examine these characterizations and discuss what LLMs can't do and what they are surprisingly good at. LLMs are still susceptible to traditional issues with AI, probabilities are not knowledge, and they are not in the world. Nonetheless, LLMs, despite not being human, have great potential to perform various creative tasks. We conclude that LLMs are beyond "mere generation" and perceivable as creative, but we may need to reassess some frameworks for creativity evaluation.

Introduction

In the past few months, popular awareness of Large Language Models (LLMs), particularly GPT-3 (Brown et al. 2020), ChatGPT, GPT-4 (OpenAI 2023), and others, has risen abruptly. The release of ChatGPT allows anyone to interact with an LLM and led to apocalyptic headlines about issues ranging from high-school essays and the future of news columnists' jobs to a massive influx of generated stories submitted to sci-fi/fantasy magazine *Clarkesworld* (Acovino, Kelly, and Abdullah 2023). However, another thread has been common: characterizing LLMs as "reasoning", "sentient" or "knowing".

Here, we investigate this kind of argument and the implications when LLMs used for creative purposes. First, we argue that these characterizations misrepresent the LLMs' behaviour: probability distributions are not minds, and the "reasoning process" of an LLM is fundamentally different from either planning agents or humans. LLMs clearly demonstrate new features and exhibit capabilities not seen before, and it is, therefore, appealing to ascribe certain properties to them and interpret their behaviour as human-like. However, given the fundamental differences, we should proceed carefully. Second, we show that Computational Creativity (CC) evaluation frameworks may need to be reassessed to accommodate the new features and behaviours of LLMs. We perform a brief creativity evaluation of LLMs using standard criteria (Ritchie 2007; Runco and Jaeger 2012), and explore if they have moved beyond "mere generation" (Ventura 2016), and if they can be perceived as creative (Colton 2008). We conclude with suggestions to further investigate LLMs as creative systems.

Background

Large Language Models appeared around 2017, and dramatically changed both CC and natural language processing. LLMs leverage transformer-based architectures (Vaswani et al. 2017) to establish a probability distribution over outputs based on properties of the word distribution in the training data, processing all tokens in an input at the same time. In particular, at each position, the influence of previous words in both the prompt and the output of the LLM can vary: in this manner, the LLM can maintain the name or gender of a character across sentences, and focus on sentences of high fluency. The degree to which much earlier words influence later words depends on the model, as does the richness of the probability distributions: models with many more parameters better maintain long-distance continuity, and allow for more subtle interactions between the words of a paragraph.

Besides model size, the volumes of training data has similarly exploded, enabling them to work with an astronomical variety of information. As a result, the probability distributions can be implicitly *conditioned* by prompt engineering: one can alter the type of response obtained by changing the rhetorical tone of the prompt (i.e. "I bet you don't know the answer to this question:"), or by giving a role or a persona in the prompt (i.e. "You are a sceptical scientist: do vampires exist?"). The ability to invoke new modes or personas (Kojima et al. 2022), allows style changes of the model in both obvious ("You are Walt Whitman; write a poem about a clam.") and less obvious ways ("You hate poetry and think it's a waste of time; please write a review of this poem:"). GPT-4 even writes code to draw images using SVG or TikZ (Bubeck et al. 2023).

Transformers can be enhanced in a variety of methods. First, it is necessary to make a clear distinction between models complimented with Reinforcement Learning from Human Feedback (RLHF), and those are not. We see clear evidence in the difference between the earlier GPT-2 and GPT-3, and ChatGPT and GPT-4: the former models are truly general, and their purpose is to create utterances in the pattern of their training distributions, while the latter operate as a chatbot, with its output probabilities tweaked, to make interacting with it more "chat-like". However, these tweaks come at a cost, as it outright refuses to write violent fiction or pornography or even discuss important political speeches or religious text. This potentially causes a substantial dent in its creative capabilities.

An alternative frame for changing the overall distributions of LLMs is to alter their training data by fine-tuning on a specific corpus of data: the model keeps its fluency while generating sentences consistent with the probability distributions of the fine-tuning data. In addition to generating poetry in a particular author's style (Sawicki et al. 2022), this approach can also yield transformers more able to correctly answer basic mathematical problems or make valid logical arguments (Cobbe et al. 2021).

What LLMs can't do

Here, we discuss several ways in which transformers do not actually reason, and why that matters for discussions of their "sentience" or other perceived properties. Key to transformers is that they sample from a probability distribution. Their structure is in this sense a very high-order Markov chain. They do not model discourse, or have "state" (besides conditioning); at best, these are implicit in the distribution.

There are legitimate questions about how a mind overall differs from a Markov model, or some other probabilistic automaton, and philosophy of mind explores the complex connections between language and consciousness. Still, human minds engage in tasks like deductive and inductive reasoning, analogic analysis, and other steps that are at best simulated by an LLM. It is seductive to assume that when an LLM estimates the probability of "True" or "False" being the right answer to a question, it is engaging in proper reasoning. However, even with curated training data, even if it can identify faulty arguments with higher probability, what is happening must be properties of the sentences analysed, perhaps in a "Clever Hans" sort of framework (Sturm 2014).

Longstanding concerns about AI also apply to transformer-based models. The most basic of these is that the AI is not an embodied agent in the physical world, but is merely a symbol-processing agent. This naturally turns into Searle's Chinese Room dilemma (Searle 1980), but the big-data version of it: does an LLM with billions of parameters still fail to "know" anything about the language operations it simulates and represents? In theory, a human or team of humans could simulate the many, many steps involved in producing a sentence from an LLM without understanding the steps to make that sentence, opening file cabinets full of topical parameters and repeatedly calculating neural network inference steps. LLMs are no different from any other artificial intelligence agent. Searle's Chinese Room dilemma may be less obviously a hindrance in a world with billions of operations per second and parallel models that store trillions of parameters: perhaps the analogy does break down.

Even though the high quality of their output may persuade otherwise: LLMs are not in the world (Dreyfus 1992), and might never be (Fjelland 2020). They cannot observe their environment apart from the training data and prompts. This is easy to demonstrate when we ask questions that require metacognition and theory of mind (Premack and Woodruff 1978). Consider asking a mental health professional the following: "I'm unhappy. What can I do to become happy again?" The patient is relying upon expertise derived across a career: the clinician must model the patient's state of mind and their previous responses to difficult situations and the clinician's therapeutic style, to find the appropriate treatment. These meta-evaluations are outside and LLM's capacities, even a model trained using RLHF with a long-term assessment of successes and failures in the model's use as a therapeutic partner would still fail at proper metacognition or modelling of patients' state. It can only consult its probability distribution and produce probable words, resulting in a generic answer about what makes *people* happy instead of what makes *the patient* specifically happy. One could change the prompt and include personal information so that the LLM gives a less generic answer about what makes *similar people* happy, but again, not what makes *the patient* happy. At best, the additional information could be viewed as a limited, volatile model of the patient.

In CC, the hype of LLMs has led to not only generating creative artefacts such as poem and story generators or other writing assistance tools, it also opens the interesting space to explore LLMs as evaluators of creative output. However, this requires a show of understanding and knowledge, especially if these evaluations are then put directly into the world. Consider an LLM that is asked to evaluate jokes (Goes et al. 2022). It is prompted with a joke (the object) and various personality descriptions (conditions), and asked if it is "funny" or "not funny". Testing a joke against multiple personalities then allows exploring how the joke works for different people and backgrounds. We identify a grounding issue with the use of LLMs as creativity evaluators. How do we know the response is meaningful? In a classification scenario as above, the tool appears to be successful, but the model predicts the next token, and not its meaning. LLMs only learn relations between words, unlike humans, who learn relations between words and the world. In other words, they lack grounding in their communication (Clark 1996).

What LLMs are surprisingly good at

One astonishing feature of LLMs is its ability to imitate the style of authors, which is a genuine creative task on its own (Brown and Jordanous 2022). It can easily, given enough training data, rewrite a few sentences in another style, including a style not attached to an individual, such as "the style of a fourth grader". Such prompts, allow the attention mechanism to shift the probability distribution to vocabulary words used by these simulated personas.

Another (perhaps not surprising) thing that LLMs can do very well is incorporating much larger amounts of information than an ordinary human can be expected to know; for example, while they may not be "reasoning", their probability distributions can incorporate philosophy papers, legal articles, medical journals and more. (Gao et al. 2020). ChatGPT can make more reasonable claims about Brazilian history than any author of this paper, as none of us knows anything about that topic. That said, LLMs may hallucinate and generate incorrect claims (OpenAI 2023). Still, on topics that rarely occur in the training data, the quality can be particularly poor (Bubeck et al. 2023).

Finally, LLMs are fantastic systems for combinational and exploratory creativity (Boden 1992). Prompting a model for variations of the same idea can appear to simulate the creative brainstorming. One can endlessly ask GPT-3 to come up with alternative uses for common objects (a standard way of testing human creativity) (Stevenson et al. 2022). Indeed, one delightful possibility is to use them in a Mad-Libs style, to fill in holes in sentences or poetic lines in surprising ways, exploring the lower-probability words in the transformer's conditional distribution. LLMs can combine styles of poets or authors and interpolate between the two. LLMs enable one to explore, mix, and match between different styles, stories, and other ideas.

LLMs and creativity

The CC community has over the years outlined several methods and standard criteria for evaluating creative systems (Ritchie 2007; Runco and Jaeger 2012). LLMs demonstrate substantial new features and behaviours that warrant an evaluation of their creativity. In particular, we explore if LLMs are beyond "mere generation" (Ventura 2016), and if they can be perceived as creative (Colton 2008). These two evaluation frames are useful to analysing the LLM as a category, and we approach the evaluation not to limited specific creative task or system.

Are LLMs beyond "mere generation"?

In general, machine learning models cannot escape their training data, and LLMs are no different, as exemplified by their unawareness of recent events. However, we can explore novelty "within the scope" of the training data.

Novelty generally occurs as a result of prompt engineering. If we asked the model to complete a prompt without any further information or context (such as inducing personalities or a specific setting) it will provide very average answers. If we ask it to *just* write a story; it produces essentially the same story. We argue that this exhibits low novelty. By providing additional information and context, we can steer and skew the output distribution in such a way that it produces results that are more novel, but the question remains: who produces the novelty? Is it the human through prompt engineering, or the machine? Given an extensive prompt, the result is not so surprising or novel. On the other hand, the "scope" of training data is so vast that LLMs can generate novel output and cause surprise to its users. *Typicality* is in a similar spot. A probability distribution, by definition, should generate typical objects. By design, LLMs produce typical objects to the training data, following the structures found in human creative output.

The outputs of LLMs are in general of high (grammatical) quality. However, from the perspective of what the output means, the quality is often poor, and often contains fabrications. This is clearly harmful when asking, for example, for medical advice as they may suggest a lethal dose (Birhane and Raji 2022). Overall, LLMs are helping people to be more productive, however, the situation with *Clarkesworld* (Acovino, Kelly, and Abdullah 2023) indicates there are some issues with scale and *value*.

Another problem for novelty, typicality, and value is the safety constraint for safety (using RLHF). These constraints limit what the LLM will generate, reducing variety and the potential for novel outputs, and increasing typicality. The quality of the output in earlier versions of GPT-4 (Bubeck et al. 2023) is very different from what you get from the version that was eventually released. This negatively influences the LLM when applied to domains that are not 'chat-like', such as poems, stories, and drawings (or the code that draws them), making the object and the system less valuable.

Besides the standard criteria, Ventura (2016) requires *intentionality* to determine if a system has moved beyond "mere generation". Intentionality is defined as; being deliberative and purposive, and the product correlates with the objective and the systems' creative process. The LLMs goal here is to generate the best possible output given the prompt given its training distribution. The most straightforward way to test intentionality, is to simply ask the LLM to explain itself, and ChatGPT often does this automatically when asked to write code. However, this ability is not that surprising, since LLMs are trained on explanations (given the large variety of Q&A websites). Moreover, these explanations are still subject to hallucinations, somewhat invalidating intentionality, and until told otherwise, the LLM will accept the hallucination as fact. Ventura's expedition ends with a generative algorithm that engages in an iterative process until it is satisfied. While the LLM is unable to engage in this process autonomously, it can clearly perform the task when given a theme by the user and asked to generate and improve a story over a few iterations. This approach is limited and only works for a few steps, but it nevertheless attempts to come up with variations and reasonable explanations.

Many CC researchers have focused on intentionality as a key area in which computers differ from humans: in this thinking, humans *choose* their activities, while computers are merely programmed to do specific things by humans. While clearly correct, the lack of agency in a chatbot or other LLMs is not an essential difference to a human, who may "choose" to answer questions, but only in the sense that capitalism requires adults to sell their labour. Intentionality also attaches in somewhat complicated ways with software, as it may also represent the intentionality of its programmers, or their bosses. In other words, their intentionality might be linked with their owners' needs and goals.

Are LLMs beyond "mere generation"? As we draw a "line in the sand", it is clear that LLMs have a good chance of producing output that is novel and valuable, and both are intentional in the sense that the LLM can reasonably explain itself and iterate on previously generated output. However, following the discourse in this paper, we find it increasingly challenging to use the evaluation frames provided by Ritchie (2007) and Ventura (2016), given the scale at which LLMs operate, how they represent and use "knowledge", and how they are made available by their owners. LLMs process massive amounts of information, but probabilities do not imply knowledge.

Can LLMs be perceived as creative?

The creative tripod focuses on the *perception* of three key aspects: skill, imagination, and appreciation (Colton 2008). LLMs lack the capability to imagine and appreciate, but they can give the appearance thereof.

LLMs might have *skill*: they demonstrate fluency, but skill goes beyond just technicalities. It also involves the ability to create an engaging narrative of unique style. While LLMs are reasonably competent at storytelling, their abilities are basic. Skill also involves separating fact from fiction to ensure logical and accurate writing.

Interestingly, the fabrications that LLMs produce do give the perception of *imagination*. In fact, the benefit of access to an enormous vocabulary makes is so that it never runs out of variations, but those are not fundamentally different or new in the sense that LLMs are not designed to do something that is different, such as using an objective function targeting novel styles (Elgammal et al. 2017). We may ask it for another variation, but that is a prompt engineering trick only maintainable in short-term memory.

LLMs can *appreciate* complex patterns, and subsequently, slice and dice the distribution in different ways–clearly a method that many appreciate. Another angle is self-appreciation. We can ask it to explain itself, or if the response contains mistakes, to revise its answer. This could be perceived as showing appreciation, self-reflection, or self-awareness. However, this is guided and directed by the user, and still just a simulation.

Perception is a tricky concept with LLMs. The power of these language models and characterizations that followed show that we can perceive them as something they are not. If LLMs can evoke this illusion, then perhaps a focus on perception for assessing their creativity is not sufficient.

Conclusion

After the release of ChatGPT, public opinion exploded with examples of both its abilities and its weaknesses. Often times, LLMs get over-qualified and claims are made that they have a true understanding of the world. We stress that mischaracterizations are potentially a problem for when and how to use LLMs and what to expect from them. Especially, how we assess the systems' intentionality becomes challenging, as it is very how hard to pin down how it structures and represents knowledge. When applying LLMs as (creative) evaluators, we encounter fundamental grounding problems. New features of LLMs easily enable the *perception* of creativity, but precisely for that reason, we need to be critical of what they *actually* do.

With this paper, we present an initial inquiry into the creativity of LLMs. Future work should address how LLMs perform in specific creative domains and roles. In particular, a full-scale creativity evaluation using SPECS (Jordanous 2012) needs to be considered to delve into the linguistic and domain-specific creativity of LLMs. Another direction to explore this kind of question is using the FACE/IDEA framework (Colton, Charnley, and Pease 2011), meant to aid development of CC systems, to look into LLM design (and human feedback) with specific creative tasks in mind.

Finally, we want to point out that for creativity evaluations of LLMs, we need a systematic approach to probing these systems. There is some value to developing a spectrum of prompts that tests different levels of creativity. In the case of the GPT series, OpenAI releases very little information about their models, and as a result, it is a particularly hard to perform scientific experiments, especially since human feedback causes their behaviours to change at a rapid pace.

Acknowledgments

The work of M.P. is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds. The work of D.B. is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

Acovino, V.; Kelly, M. L.; and Abdullah, H. 2023. A sci-fi magazine has cut off submissions after a flood of AI-generated stories. *NPR*. Retrieved from https://www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence (Accessed 24/04/2023).

Birhane, A., and Raji, D. 2022. ChatGPT, Galactica, and the progress trap. *Wired*. Retrieved from https://www.wired.com/story/large-language-models-critique (Accessed 24/04/2023).

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Brown, D. G., and Jordanous, A. 2022. Is style reproduction a computational creativity task? In *Proceedings of 13th International Conference on Computational Creativity*, 220–229.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Clark, H. H. 1996. *Using language*. Cambridge: Cambridge University Press.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Colton, S.; Charnley, J. W.; and Pease, A. 2011. Computational creativity theory: The face and idea descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95. Mexico City.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *2008 AAAI Spring Symposium*, 14–20.

Dreyfus, H. L. 1992. *What computers still can't do: A critique of artificial reason*. MIT press.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. In *Proceedings of the 8th International Conference on Computational Creativity*, 96–103.

Fjelland, R. 2020. Why general artificial intelligence will not be realized. *Humanit. Soc. Sci. Commun.* 7(1):1–9.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd Score: A method for the evaluation of jokes using large language model ai voters as judges. *arXiv preprint arXiv:2212.11214*.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4:246–279.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Premack, D., and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1(4):515–526.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds & Machines* 17:76–99.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Sawicki, P.; Grzes, M.; Jordanous, A.; Brown, D.; and Peeperkorn, M. 2022. Training GPT-2 to represent two romantic-era authors: Challenges, evaluations and pitfalls. In *Proceedings of 13th International Conference on Computational Creativity*.

Searle, J. R. 1980. Minds, brains, and programs. *Behav. Brain Sci.* 3(3):417–424.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*, 164–168.

Sturm, B. L. 2014. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia* 16(6):1636–1644.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24.

The Spectrum of Unpredictability and its Relation to Creative Autonomy

Iikka Hauhio, **Anna Kantosalo**, **Simo Linkola**, and **Hannu Toivonen**
Department of Computer Science
University of Helsinki, Finland
{first.last}@helsinki.fi

Abstract

Recent popularity of generative AI tools has sparked discussion on how the unpredictability of the tools affects the creativity of the human and the AI program alike, as unpredictability prevents the human user from fully controlling the output. We present a framework for categorizing unpredictability on four different dimensions and analyze the types of unpredictability found in generative AI tools. We also describe the relationship between unpredictability, uncontrollability, and Jennings' creative autonomy. We conclude that while unpredictability does not on its own imply creative autonomy, it could be used as a central condition for it, if accompanied by other conditions.

Introduction

The recent popularity of generative and creative artificial intelligence (AI) has raised the relationship between AI and creativity to common debate. For example, in a recent decision, the United States Copyright Office determined that in certain situations, the user of an AI image generation tool is not considered the author of the work for copyright purposes, because the tool works in an unpredictable manner.[1]

Many of today's generative AI systems are unpredictable in various respects and to varying degrees. If the unpredictability of the system rules out the user's (full) authorship of the generated results, who or what can be attributed with creativity when the end result itself is considered creative, i.e., novel and valuable (Runco and Jaeger 2012)? Is it a reasonable argument that the program must in that case have committed creative acts?

In this paper, we present a categorization for unpredictability that can be used to analyze different scenarios in which unpredictability can affect the creativity of the system. We argue that unpredictability may help to characterize the creative autonomy of the system, defined as "the system's freedom to pursue a course independent of its programmer's or operator's intentions" (Jennings 2010). Unpredictability implies that the human user does not have complete control over the system, which is a requirement for creative autonomy of the system.

Throughout this paper, we assume that one is assessing a creative system. We use language models and image generators as example tools without making claims about the creativity of any specific tools for any specific tasks. Rather, the arguments we present are philosophical in nature, asking the following question: assuming that the outputs of a system are creative, how does its possible unpredictability affect our judgement of the creative role and autonomy of the system?

The rest of the paper is organized as follows. We first present a definition for unpredictability and categorize different types of unpredictability. We then analyze unpredictability of concurrent generative AI programs. Finally, we present an argument that connects unpredictability to uncontrollability and thus Jennings' creative autonomy (Jennings 2010).

Unpredictability

In this paper, we define *unpredictability* as the inability of an *observer* (e.g. the creator of a generative program or its user) to determine the generative *outcome of a program* given a specific *input*. The observer can be seen as an entity that holds a certain amount of information about the program through knowledge of its internal workings or holistic observation of the program at work. This suggests that unpredictability from the point of view of a user may be influenced by experience and thus has an element of time to it and links it to other experiential properties of generative systems, such as surprise (e.g. (Grace et al. 2015)). This makes unpredictability a meaningful concept for evaluation of computationally creative systems that allows us to compare systems and link it to the discussion of meaningfully assigning autonomy to an AI.

Unpredictability is not a characteristic that uniformly covers the whole output of a system. Rather, it is a question of perspective. If the user prompts a language model to produce a poem, it usually is predictable that the output is, or resembles, a poem, while many details about the structure and word choices might be unpredictable. In the case of generative systems, it is important to define the extent of unpredictability; in this paper, we call the (unpredictable) features of interest the *outcome* of the system. The outcomes

[1] "Rather than a tool that [the user] controlled and guided to reach her desired image, Midjourney generates images in an unpredictable way. Accordingly, Midjourney users are not the "authors" for copyright purposes of the images the technology generates." (Kasunic 2023)

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

in our case are features of the artifacts the tools produce, i.e., features that the user might want to control but cannot due to unpredictability. These features can be concrete, such as the exact colors used by an AI image generator, or more abstract, such as the mood expressed by a generated poem.

We define unpredictability with respect to the process, the outcome, and the observer, and we will similarly categorize the different types of unpredictability based these three dimensions: (1) the *cause* of the unpredictability, i.e., what kind of process causes the outcome to be unpredictable; (2) the *scope* of the unpredictability, i.e., what types of outcomes are unpredictable; and lastly (3) the *point-of-view* of the unpredictability, i.e., who is the observer that determines that the process is unpredictable. We also consider (4) the *duration* of the unpredictability, i.e., when the predictions are performed.

Causes of Unpredictability

We divide unpredictability to three categories based on the cause of unpredictability: stochastic (indeterministic), chaotic (deterministic), and mixed-cause unpredictability.

Stochastic unpredictability refers to indeterministic unpredictability that cannot be predicted by the observer. It is similar to Boden's *absolute unpredictability* (Boden 2004); however, our definition includes technically deterministic processes which cannot in practice be predicted, such as pseudo-random number generators initialized with unknown, randomized seeds. From the point of view of the observer, the processes in this category are random. If the outcome of a program is stochastically unpredictable, the outcome will change unpredictably each time the program is run.

Stochastic unpredictability can be more strong in some scenarios than others. Compare, for example, a fair dice roll and a weighted dice. Both of them contain some unpredictability: we cannot be completely sure what the result will be. However, in the latter case, one outcome is more likely than the others. In an extreme case, a weighted dice will almost always produce the same result, thus making it completely predictable. Thus, depending on the probability distribution, some cases of stochastic unpredictability are more unpredictable than others. The exact categorization of the subtypes of stochastic unpredictability is not in the scope of this paper.

Chaotic unpredictability refers to deterministic but chaotic processes. If a program is chaotically unpredictable, its output will change unpredictably each time it is run with a new input, but it will consistently provide the same output for the same input. This category includes pseudo-random numbers generated with a known seed. Neural networks that are too complex for humans to understand (cf. Burrell 2016) might also belong to this category. In Boden's terms, this type of unpredictability is called *butterfly unpredictability* (Boden 2004).

Mixed-cause unpredictability is a combination of both stochastic and chaotic unpredictability. In practice, many generative AI programs include both types. For example, a language model-based generator might first calculate the probability distribution for the next word using a complex neural network (chaotic unpredictability), and then sample a word from the distribution (stochastic unpredictability). If the outcome is the result of both stochastic and chaotic unpredictability, it can change to some degree each time the program is run with the same input while still retaining some properties between the outcomes.

Scope of Unpredictability

We call the "size" of the set of features of the output affected by the unpredictability the *scope* of the unpredictability. Next, we sketch different levels of unpredictability based on their scope. Note that the levels are not based on shallow, technical distances such as edit distance, but rather on their semantic distance. Here, we outline the idea, and a more exact characterization is left for future work.

Low-level unpredictability occurs when the unpredictable variation affects small details or minor choices in the output, e.g., the exact word choices of a poem generator or the exact colors produced by an image synthesis model cannot be predicted.

Middle-level unpredictability refers to unpredictability of broad details and major choices in the output. In a poem generator's output, this might mean features such as the symbols used, or the meter followed. In an image synthesis program, middle-level features might be the objects included in the scene, the layout of the image and the art style used.

High-level unpredictability refers to even more abstract features, such as the topics included in the work. At the highest level, even the artifact class itself could be unpredictable.

Point-of-view of Unpredictability

Unpredictability is defined with respect to an observer for whom the process is unpredictable. We propose the following categorization to world- and user-unpredictability, which can be compared to Boden's categorization of creativity to H-creativity and P-creativity (Boden 2004). Boden argues that if the purpose is to evaluate the capability of an individual — or a program — to be creative, then P-creativity and what we call user-unpredictability are more interesting concepts than H-creativity and world-unpredictability.

World-unpredictability refers to the situation in which no one can predict the outcome of the process. By definition, this includes all stochastic programs, but it might also include some chaotic programs if they are sufficiently complex for any human to understand (cf. Burrell 2016).

User-unpredictability refers to the situation in which the humans who choose the input to the program cannot predict the output. While weaker than world-unpredictability, user-unpredictability is still enough to establish the control the user has over the output. If the user cannot predict the outcome of the process, they cannot reliably control it (see the chapter below for elaboration of unpredictability and control).

Akin to user-unpredictability, it is also possible to define concepts such as *programmer-unpredictability* and *audience-unpredictability*, if needed.

Changes in Subjective Unpredictability

In addition to the *how*, *what*, and *who* of the previous categorizations, we can also ask *when* the program is unpredictable for a particular observer.

Permanent unpredictability lasts forever. By definition, this includes all stochastic unpredictability, but some sufficiently complex chaotic processes might also belong to this category, at least if we only consider humans as possible observers.

Temporary unpredictability can be overcome, causing the process to become predictable in time. For example, a deterministic program becomes predictable for a certain input after the first time it is run, as all the subsequent runs will produce the same result. Likewise, a process presumed to be chaotic can become predictable after it is understood better.

Unpredictability in Generative AI Programs

Large neural networks used for generative tasks, such as GPT-3 (Radford et al. 2019) and Stable Diffusion (Rombach et al. 2022), contain billions of parameters and are often regarded as black boxes due to their unexplainability. Their intrinsic complexity makes it impossible for a human to fully comprehend their operation (Burrell 2016), which implies they contain unpredictability.

We argue that following our categorization of unpredictability, most concurrent AI tools contain *mixed-cause*, *low/middle-level user*-unpredictability. Complex AI models behave both predictably and unpredictability (Ganguli et al. 2022) and contain both chaotic parts (such as deterministic neural networks) and stochastic parts (such as token-sampling in language models and random noise in image synthesis models). Since state-of-the-art models are broadly speaking quite good at following instructions specified in the prompt (Radford et al. 2019; Rombach et al. 2022), they are predictable and controllable at high-level, but not necessary at low- and middle-levels. While they contain some world-unpredictable parts (such as those which are completely stochastic), they also contain parts which become more predictable as the user gains intuition over the model's behavior, causing the model to be more unpredictable to some users than others.

$$P \ni x \overset{\text{predict } f}{\dashrightarrow} Q \ni f(x)$$

$$Q \ni f(x) \overset{\text{control } f}{\dashrightarrow} P \ni x$$

Figure 1: The difference between predicting (determining the outcome caused by the input property) and controlling (determining the input property that causes the desired outcome).

Unpredictability and Uncontrollability

To control a generative program, the user must be able to predict the correct *inputs* P that will cause the desired *outcome* Q. For example, if the user wants an image synthesis program to use a certain color (e.g., Q is the set of images with red color), they must determine which prompts cause that color to be generated (e.g., P is the set of prompts that contain the string "red"). The ability to control the program corresponds thus to the ability to calculate or estimate the inverse of the program.

Predictability, on the other hand, is about determining the outcome Q given an input P. Logically, the ability to control and the ability to predict are inverses of each other, and separate from each other. The difference is explained in mathematical notation in Figure 1. However, we argue that in practice, unpredictability implies uncontrollability.

If the program was controllable but unpredictable, it would mean that its inverse is predictable. If the program was chaotically unpredictable, its inverse would not be chaotic. If the program was stochastically unpredictable, its inverse would not be stochastic. We argue that this kind of situation is rare in the context and generative AI and creative programs overall, but we'll leave the proof for future work. In the rest of the paper, we assume that unpredictability does imply uncontrollability.

Note that the reverse is not true: a predictable program can be uncontrollable. For example, a program that always produces the same output is predictable and uncontrollable: for any desired outcome Q which is not the outcome the program produces, there exist no solutions for P.

Creative Autonomy

As discussed above, unpredictability makes it impossible for the user to completely control the AI tool's operation. Unpredictability is therefore related to Jennings' *creative autonomy* (Jennings 2010): What appears as unpredictable behaviour to the user might be explained by creative autonomy of the system. We seek to use unpredictability as a tool to characterize the creative autonomy of systems, especially of black box AI generators.

Jennings gives three criteria for creative autonomy: *autonomous evaluation*, *autonomous change of standards*, and *non-randomness* (Jennings 2010). Autonomous evaluation allows the system to observe the quality of its own work and thereby improve its operations. Autonomous change,

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

in turn, allows the system to adjust its own standards and goals. Autonomous evaluation and change could be trivially achieved with random behavior, but the third criterion rules out fully random behaviour. Jennings explicitly allows for randomness in the processes, and many creative systems have stochastic components — they just shouldn't be fully random.

The relationship between unpredictability and creative autonomy is not one-to-one. Not all unpredictability implies creative autonomy: fully random behaviour would be unpredictable but not autonomously creative. Also, not all unpredictable generative behavior is creative. Vice versa, it can be argued that some predictable processes do have creative autonomy despite their predictability, since being autonomous does not entail being unpredictable. For example, many human artists have a very constant style or paradigm they follow, while retaining creative autonomy.

It is clear that deterministic unpredictability does not necessarily entail creative autonomy, either. Consider fractal images such as the Mandelbrot set image. These images are deterministic but chaotic, and it is very hard to predict what a certain "deeply zoomed" region of the image looks like without solving the equation for the points in that region. However, they are also completely static, and no evaluation or change occurs when calculating the equation. The fractal equation does thus not have creative autonomy, although it can potentially produce novel and valuable images when solved for yet unvisited regions of the coordinate plane.

Despite unpredictability not directly implying creative autonomy, we argue that unpredictability could be used as a *condition* for it in a yet-to-be formulated framework for evaluating unpredictable programs: if the evaluations and changes that occur during the program's execution are unpredictable, they cannot be controlled by the user and are thus autonomous, assuming they are not fully random, i.e., the unpredictability should not be only stochastic.

Unpredictability could be used to show that the user is incapable of controlling a creative program, in order to provide arguments for the program's creative autonomy. To prove this for a single user during their use of the program, the type of unpredictability used as a condition for creative autonomy can be temporary user-unpredictability instead of stronger forms of unpredictability such as permanent world-unpredictability, although this would make the perception of creative autonomy subjective allow it to change over time. We leave the debate of whether this is acceptable or not and how unpredictability shapes the artist's perception of their own role and agency to further research.

Conclusions

Unpredictability is an important property of many generative AI programs and has implications to their creativity, since it limits the ability of the user to control the operation of the AI programs. We presented a framework for categorizing different types of unpredictability based on the *how*, *what*, *who*, and *when*: the causes, scopes, observers, and the change of subjective unpredictability. These categorizations can be used to characterize generative AI tools.

We discussed the relationship between unpredictability, uncontrollability, and creative autonomy (Jennings 2010). Unpredictability implies uncontrollability, which is a requirement for creative autonomy. While unpredictability does not imply creative autonomy, it could be used as a condition in a larger framework intended for determining and analyzing creative autonomy in generative AI programs. Further research should be conducted to determine a sufficient set of additional conditions to be used alongside unpredictability.

Unpredictability of complex generative systems, and the lack of control it implies, shows that it can be difficult to attribute creativity to one party only, be it the user, the developer, or the system. While the US Copyright Office's decision to deny authorship of the human who used an AI image synthesis tool is probably justified, this does not mean that the tool was the author. We argue that in this case, there simply is no single author. This does not mean, however, that there is no creativity in the process: the creativity is just not controlled by any one stakeholder. This implies that, although not necessarily autonomously creative, unpredictable programs do nevertheless play a significant role in the creative process.

Author Contributions

The idea for the paper came from IH and it was mostly written by IH. AK, SL and HT supported the concept development and edited the paper.

Acknowledgments

IH is funded by the doctoral programme in Computer Science at the University of Helsinki. SL is funded by Academy of Finland grant 328729 (CACDAR). AK works at Siili Solutions Oyj; this paper was not a part of a work assignment.

References

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3(1).

Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; et al. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3-4):125–147.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Kasunic, R. J. 2023. Re: Zarya of the Dawn (Registration VAu001480196).

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9. Viitattu: 16.6.2022.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal* 24(1):92–96.

Many Meanings of Intentionality: A Brief Disambiguation for Computational Creativity

Simo Linkola and **Iikka Hauhio**
Department of Computer Science
University of Helsinki, Finland
simo.linkola@helsinki.fi, iikka.hauhio@helsinki.fi

Abstract

Computational creativity is a multi-disciplinary field with authors from diverse backgrounds. This raises a threat of misunderstanding when the authors from different backgrounds use the same words with different meanings. We elaborate on the two main meanings of the word "intentionality" found in the computational creativity literature: aboutness and goal directness. Both of the meanings are prominent concepts in computational creativity, but they relate to very different ideas, perspectives and contexts. Aboutness, in philosophy, is quality of the mind to be about somethings outside of it, while goal directness, when interpreted from engineering perspective, is the system's property to be able to produce outputs that are aligned with its goals. We briefly explain both of the meanings, highlight their related concepts, and provide a discussion of how these two interpretations of the same word are related.

Introduction

Creativity research entails many related concepts and terms that are argued to be relevant for creativity. One of these terms is "intentionality", which has even been considered as a necessary requirement for computational creativity (CC) from a process perspective (Ventura 2016; 2017) and appended to the artefact (or concept) requirements of the standard definition of creativity including novelty and value (Boden 2004; Runco and Jaeger 2012). Unfortunately, the word "intentionality" has multiple homonymous meanings which are seemingly completely separate. This can cause unnecessary confusion in the field when researchers using the word – or referring to others using the word – do not specify exactly which meaning they intend. In this short paper, we briefly explain (some of) these different meanings and their implications, and suggest that the field, in general, takes care when using the word – or select more fitting terms which have no similar tendency to cause confusion.

Computational Creativity is a multi-disciplinary field with contributing members from different fields such as computer science, psychology, fine arts and philosophy. Because of this multi-disciplinary approach, the field is able to incorporate many perspectives to the multi-faceted phenomenon of creativity, which has been a discussion subject within the field in its own right. For example, Pérez Y Pérez (2018) describes a CC-continuum which spans from engineering-mathematical approach to cognitive-social approach, where the ends of the continuum follow different paradigms and may have different research methods as well as end goals. This produces internal tension, typical to many multi-disciplinary fields, and the field has to constantly battle with misinterpretations of the conveyed messages as the vocabulary between the members of the field may vary greatly.

In this paper, we focus on the meaning of a single word, "intentionality", because of its central status within the field of computational creativity. In philosophy, also adopted to cognitive science, intentionality means "aboutness", the ability of the mental states to refer to objects outside the mind (Schlicht and Starzak 2021). In other contexts, intentionality may refer simply to the property of having intentions, goals, or objectives (Ventura 2016; 2017). At a first look, these meanings have nothing to do with each other but sharing the same name, yet the word is used without referring to its explicit meaning in the existing computational creativity literature (Colton and Ventura 2014; Grace and Maher 2015; Varshney 2020; Sewell, Christiansen, and Bodily 2020).

The rest of the paper is organised as follows. Next, we provide a brief disambiguation of the two main meanings of the word "intentionality" especially relevant for the field of computational creativity. Then, we address how the different meanings may possibly relate to each other. We end the paper with conclusions.

Interpretations of Intentionality

In this section, we review the different meanings of the word "intentionality" relevant for computational creativity. We begin with the philosophical concept of *aboutness* and follow with more engineering-oriented *goal directness*.

Aboutness

In philosophy, intentionality refers to the quality of mental states to be "about" something or "directed to" something (Schlicht and Starzak 2021), popularised by Franz Brentano, who argues that it is what separates mental states from physical states (Brentano 1874). While intentionality is a quality of mental states, it applies to words and symbols as well if they are being processed by a mind. That is, an

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

intentional mind can produce meaningful words and symbols, and as well interpret words and symbols as having a meaning.

Intentionality of computer programs is a debated topic. Searle (1980) argues with his famous Chinese Room thought experiment that computers – and computer programs – are not intentional, and intentionality is exclusively a feature of biological brains and "equivalent" systems. Following Searle's thought experiment, Harnad (1990) formalised the question of how symbolic computation systems could become intentional into the so-called *symbol grounding problem*: how can symbols have meaning, if they are only defined in terms of other symbols in the computation system? Harnad argues that in order for the system to be intentional (i.e., its symbols be grounded), they must have intrinsic, not extrinsic meaning. He calls extrinsic meanings *parasitic*, since they rely on outside observers for interpretation. That is, if the program is not intentional, its states are meaningless by themselves. Conversely, the states of an intentional program have self-standing meanings.

Intentionality is an overarching property of all mental states, and is thus not only applied to communication. However, in the context of generative artificial intelligence, the focus is most often in the inputs and outputs of the program. When analysing intentionality pertaining to the output, the term *communicative intent* (Bender and Koller 2020) or *authorial intent* (Barten 1967) is often used to refer to the meaning of the output as "intended" by its producer. This term has some overlap with the term "goal" as in "goal directness": the communication is typically done for a purpose, and thus the communicative intent is often directly related to a goal of a program. However, not all communicative intents are necessarily tied to a goal: for example, a person might accidentally say something that is against their goals, in which case the words still have meaning, but they are not aligned with the goals of the person. Another example might be a person speaking their stream-of-consciousness, saying what comes to their mind. In both of these cases, the speech act as a whole might still have a goal, but the individual parts of the act might not.

Intentionality can be seen as a crucial part of both creativity and cognition in general. It enables understanding (i.e., retrieving meaning of) mental states, including one's own processes, goals, memories, and so on, but also interpreting and determining value of texts and other artefacts. If a program is not intentional, its creativity and the value of its works are parasitic, dependent on outside observers.

However, not all consider intentionality important. Most famously, Roland Barten argues in his essay "The Death of the Author" that authorial intent should play no role in the interpretation of a text (Barten 1967). According to his view, literary works should be regarded as eternal objects, discovered rather than created, that have "no origin but the language itself". Under this kind of viewpoint, the communicative *intent* of the author no longer matters, but merely the communicative *function* of the words, i.e., how they are actually interpreted: if it produces something novel and valuable for the observer, it doesn't matter how it did it.

Goal Directness

In the field of computational creativity, Ventura (2016; 2017) defines intentionality as

> the fact of being deliberative or purposive; that is, the output of the system is the result of the system having a goal or objective – the system's product is correlated with its process.

Ventura (2016; 2017) does not provide an explanation for his definition to let us better understand the influences behind it; we interpret this to mean that he assumes the definition to be general enough – perhaps in line with the everyday usage of the word – that it can be accepted without explicit references. Thus, next, we provide a brief disambiguation of the definition to explain how we interpret the definition and what other possible – though mostly improbable – interpretations there could be.

The above textbook definition conflates the property of (1) being deliberative or purposive, with the notions that (2) the output of the system is the result of the system having a goal and (3) the output of the system correlates with the process of the system. All three statements of the definition are used to illustrate what intentionality in general terms means without having to construct a more formal definition, yet each of them alone may be interpreted to mean something subtly different.

Being deliberative or purposive may refer to a human (or animal) capability. That is, this statement alone can be interpreted to imply that the entity with this property has similar mental state quality as "aboutness". On the other hand, it may imply that there are other properties, such as the next two statements which are used in the definition to clarify what the first statement means.

The output of the system is a result of the system having a goal can be interpreted at least in two different ways: (1) the system has a goal to produce (certain kind of) outputs in general or (2) the production of a single output is affected by a goal. The main difference between these two interpretations is the time scale: on (1) the focus is on the general functionality and goals of the system while on (2) the focus is on a single artefact production process. The second interpretation can be further elaborated as (2a) the production of the output began because of a particular goal or (2b) the output aims to represent a particular goal of the system.

The system's product is correlated with its process can be interpreted in as many ways as an output can be correlated with the process it was produced by. One straightforward interpretation is that the process producing the output varies – somehow meaningfully or within reason – based on what the outputs aim to represent. That is, the process varies based on the goals for this particular output.

The most likely interpretation of the last two statements forms a description of a system producing outputs which aim to represent particular goals (2b, above), and how the system produces the particular outputs is affected by the goals aimed to be represented or fulfilled by the outputs. In other words, *the goals of the system (for particular outputs) affect the production process and, thus, the outputs*.

The above interpretation of Ventura's intentionality definition is in line with the concept of *goal directness*, a property where the behaviour of the system is aimed towards a goal or a completion of a task. This interpretation does not state anything about how this behaviour is achieved, yet if the first statement of Ventura's definition is also taken into account, it may be implicitly assumed that the system has human characteristics such as "aboutness".

Unfortunately, goal directness is another term subject to misinterpretation as its origins are in psychology (Frese and Sabini 2021). However, the term is frequently used in the context of artificial intelligence and intelligent agents loosely in the same way as the last two statements of Ventura's intentionality definition. Goal-based and utility-based agents (Russell and Norvig 2010) both have the capability of goal-directed behaviour, while learning agents can do so in adaptive deployment environments.

Overall, goal directness is associated with many other terms and concepts, e.g., in philosophy, artificial intelligence, and software engineering.

First, in software engineering, self-adaptive and self-aware systems (Kounev et al. 2017) aim to account for more appropriately exhibited goal-directed behaviour by allowing the system to change how it operates towards its goals based on the observed context and the system state. Linkola et al. (2017) elaborate on the concept of self-awareness in the context of software architectures for artificial, creative systems. By their argumentation, goals are one of the most notable aspects for the creative systems to be self-aware of.

Second, the question of from where the goal-directed behaviour originates, i.e., what *motivates* it and who selects the goals, is interesting. On the philosophical side of computational creativity, Guckelsberger, Salge, and Colton (2017) have studied the notion of *why* a creative system does what it does, arriving to the conclusion that, in the end, it is nearly always the programmer who decided the goals. Furthermore, motivation of intelligent agents in general has been discussed, e.g., in the context of reinforcement learning (Schmidhuber 2010; Barto 2013).

Third, the concept of creative autonomy (Jennings 2010) is related to not only goal-directed behaviour but also to motivation and self-awareness. To have creative autonomy, the system must fulfil three requirements: autonomous evaluation, autonomous change and non-randomness. Autonomous evaluation states that the system must be able to evaluate its own outputs. Thus, it directly relates to goal directness, as being able to evaluate what the system itself did, in many cases, provides ways for the system to reach its goals better. Autonomous change states that the system must be able to change its own evaluation standards, which relates to the meta-level discussion of goals and their adjustments, a prominent focus in self-adaptive and self-aware systems. Lastly, the non-randomness states that the system's evaluation or change of evaluation standards is not purely random. This statement relates to the process part of the goal-directed behaviour. While it does not state which processes should be used or how much non-randomness there should be, it is argued that concepts of self-adaptive and self-aware systems as well as motivation can help in satisfying this requirement.

Bridging the Gap

At first sight, when comparing the philosophical concept of aboutness and the engineering interpretation of goal directness, it may seem that these two concepts have little in common. However, their relationship becomes more evident if we take a look of the terms related to goal directness in the field of psychology.

In early behaviourism (Frese and Sabini 2021), to avoid long-winded discussion, goal-directed behaviour was associated with the problems of *teleology*, the idea that the future acts upon the past, and *prevision*, a plausible account of the anticipation of a goal. Prevision can be further explained, e.g., with *representations* by which an organism can evaluate the results of their behaviour, but the dichotomy of having a purpose and how to act on that purpose is not resolved. Nonetheless, an organism must have aboutness in order to have representations of the results of the behaviour to associate the mental representations with the real world.

In more modern psychology (Frese and Sabini 2021), the notion of the negative feedback between the goal state and the current state avoids the teleological conundrum; the goals of the behaviour can affect the behaviour which is aimed at fulfilling those particular goals. This feedback loop, as an abstract notion, does not require mental states to have any particular quality, and it can be used to provide structure to the behaviour, e.g., long-term planning. The very idea of feedback loop is also a basic building block, in various forms, in modern intelligent agents (Russell and Norvig 2010; Barto 2013), and thus the usage of the term goal-directed behaviour is apt for intelligent agents from this perspective.

However, there are still some nuances that are not captured with the above elaboration of goal-directed behaviour when relating it to aboutness. Aboutness is a quality of the mental states; a mental state is about something and that something can be a cat, a building plan, the objective function of an AI program, or the communicative intent we assume another person to have. That is, aboutness is an overarching quality of the mind with plethora of application targets depending on the context and priming. Goal directness, on the other hand, does not specify whether the system can interpret a single type of goal or multiple types of goal, and how abstract these goal types are. That is, to even begin to argue that an artificial system has some notion of human aboutness in itself, it should be able to reason about vast number of different concepts on different abstraction levels – potentially changing the reasoning domain or process during the procedure. This kind of behaviour is not often covered or measured by the concept of goal directness in practical artificial intelligence or computational creativity.

Aboutness as a quality of mind that can manifest reflection for different phenomena, not only for the behaviour of oneself, brings about further differences. For example, aboutness can be present in elaborating the communicative intent the other person has when conveying a message. This kind of reflection, which is not directly about the environment and the agent's goals, is not always covered by AI techniques which can be deemed to satisfy some level of goal directness. While the AI field also tackles these problems, the fact that they require different solutions implies that the aboutness, as understood in philosophy, is a more general phenomenon than what AI techniques currently cover.

In artificial systems, both of the above differences of functionality provided by aboutness with respect to functionality assumed by goal directness, multiple application targets and multiple abstraction levels, can be tackled to some extent with meta-reasoning and other meta-level approaches such as self-adaptive and self-aware systems. However, the main philosophical debate still exists: whether a computational system can have aboutness as its own quality or not.

Conclusions

In this paper we have discussed the meaning of the word "intentionality" in the context of computational creativity. We have explained the two main interpretations, aboutness and goal directness, and provided a brief analysis on how these interpretations can be aligned with respect to each other. The main point of this disambiguation was to show how these interpretations are different and what concepts or qualities of the interpretations are related.

Overall, we propound authors of computational creativity papers to be aware of these different interpretations and, should they use the term in their papers, clearly state which interpretation they are referring to. Using the same word with different interpretations – some of which are conceptually more challenging to replicate in machines – may cause dilution of the multi-faceted and nuanced philosophical concepts into luke-warm engineering solutions.

Acknowledgments

SL is funded by Academy of Finland grant 328729 (CACDAR). IH is funded by the doctoral programme in Computer Science at the University of Helsinki. The authors wish to thank Anna-Mari Rusanen for her advice on the topic of intentionality in philosophy.

References

Barten, R. 1967. The Death of the Author. *Aspen*.

Barto, A. G. 2013. Intrinsic motivation and reinforcement learning. *Intrinsically motivated learning in natural and artificial systems* 17–47.

Bender, E. M., and Koller, A. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185–5198.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London, UK: Routledge, 2nd edition.

Brentano, F. 1874. Psychology from an empirical standpoint (trans. L. MacAlister). *New York 1973*.

Colton, S., and Ventura, D. 2014. You can't know my mind: A festival of computational creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*, 351–354.

Frese, M., and Sabini, J. 2021. Goal directed behavior: The concept of action in psychology.

Grace, K., and Maher, M. L. 2015. Specific curiosity as a cause and consequence of transformational creativity. In *Proceedings of the Sixth International Conference on Computational Creativity*, 260–267.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the" why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency.

Harnad, S. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena* 42(1-3):335–346.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Kounev, S.; Kephart, J.; Milenkoski, A.; and Zhu, X. 2017. *Self-Aware Computing Systems*. Springer International Publishing.

Linkola, S.; Kantosalo, A.; Männistö, T.; and Toivonen, H. 2017. Aspects of self-awareness: An anatomy of metacreative systems. In *Proceedings of the Eight International Conference on Computational Creativity*, 189–196. Atlanta, Georgia, USA: Georgia Institute of Technology.

Pérez Y Pérez, R. 2018. The computational creativity continuum. In *Proceedings of the Ninth International Conference on Computational Creativity*, 177–184. Salamanca, Spain: Association for Computational Creativity.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity research journal* 24(1):92–96.

Russell, S. J., and Norvig, P. 2010. *Artificial intelligence: A modern approach*. Pearson Education, Inc.

Schlicht, T., and Starzak, T. 2021. Prospects of enactivist approaches to intentionality and cognition. *Synthese* 198(Suppl 1):89–113.

Schmidhuber, J. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2(3):230–247.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3(3):417–424.

Sewell, A.; Christiansen, A.; and Bodily, P. M. 2020. Creative constellation generation: A system description. In *Proceedings of the 11th International Conference on Computational Creativity*, 496–499.

Varshney, L. R. 2020. Limits theorems for creativity with intentionality. In *Proceedings of the 11th International Conference on Computational Creativity*, 390–393.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the Seventh International Conference on Computational Creativity*, 17–24. Paris, France: Sony CSL.

Ventura, D. 2017. How to build a CC system. In *Eighth International Conference on Computational Creativity*, 253–260. Atlanta, GA, USA: Association for Computational Creativity.

Should we have seen the coming storm? Transformers, society, and CC

Carolyn E. Lamb[1] and Daniel G. Brown[2]
[1] School of Computing, Queen's University, Kingston, Ontario, Canada
[2] Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
cel4@queensu.ca, dan.brown@uwaterloo.ca

Abstract

Since ICCC 2022, transformer models easily usable through natural language prompts have changed the face of computational creativity. They raise disquieting social and legal issues. The ICCC community was to a great extent unprepared. We review some harms, dangers, and questions raised by transformer models and recommend that the CC community must more widely and urgently attend to the social impacts of CC.

Introduction

Large Language Models (LLMs), particularly based on transformers (Vaswani et al. 2017), have shown increasing success at generating text and other forms of media, such as visual art, based on natural language prompts. This success has brought us to a tipping point in the social role of computational creativity (CC). Suddenly, our field of research is no longer an obscure curiosity, but a major news item.

This increased prominence comes with increased ethical scrutiny and fear of harm. In fact, the release of high quality transformers that can be used by a human without any programming ability, like Midjourney and ChatGPT, has already significantly affected human artists. The level of ethical, societal, and legal concern caused by these new systems appears to have taken our community by surprise. While some may draw a technical distinction between CC and generative AI, the two fields are sufficiently similar that public opinion and social impact are not likely to distinguish between them.

We argue that the CC community must pay urgent attention to LLMs' ethical issues. Otherwise we risk losing relevance as public perceptions and concerns regarding CC systems shift out from under us. Transformers risk the "malevolent creativity" described by Cropley et al (2008), where a system's output is novel and valuable to *someone*, but harmful to society at large.

We summarize a few of these ethical issues and perform a brief literature review showing that, while the CC community has been increasingly aware of creative possibilities of LLMs, analysis of their societal and legal effects has lagged. We finish with recommendations for how the CC community should address these issues.

Transformer models and society

Ethics of LLMs have been discussed since their inception. Bender et al (2021) provide an important summary. Unintended bias, toxicity in training sets, or deliberate misuse of these models all result in harmful output. The energy required for training and using these large models causes environmental harm. Below, we highlight some further issues relevant to CC.

Plagiarism and human replacement

The enormous training sets used by transformers contain copyrighted text and artwork by humans, collected by web crawling without consent (Zirpoli 2023). They can generate an unlimited amount of new work, including text or art "in the style of" a particular human, which can be used in place of the human's work. Art made by these models has already been used in contexts where traditionally a human would be employed (Schaub 2022). Generative AI may thus enable corporate groups to produce endless content for audiences without compensating rightsholders, destroying prospects for humans in creative careers (Sobel 2017).

In the US, "fair use" protects use of copyrighted materials for certain purposes, and might protect AI training (Lemley and Casey 2020), though other countries' laws may differ (Brown, Byl, and Grossman 2021). Sobel's (2017) review suggests a double bind: if AI training is not fair use, scientific progress is hindered, but if AI training *is* fair use, writers and artists suffer. The fair use claim is being tested in court via lawsuits against Midjourney and Stable Diffusion (Zirpoli 2023) for training on living artists' copyrighted work without consent. Getty Images has also sued Stability AI for training Stable Diffusion on Getty's photos without a license, sometimes producing output so close to training images that it contained the same watermark (Belanger 2023).

Professional writers also are concerned. For instance, the science fiction magazine Clarkesworld recently closed submissions due to a rush of AI-generated submissions (Xiang 2023b). Hundreds of low-quality AI-generated books have also appeared in Amazon's Kindle store (Bensinger 2023). Since Kindle Unlimited distributes proceeds between all participating authors, these AI authors are siphoning income from human authors (Scalzi 2023). In journalism, the use of AI to replace humans is also increasing (Sweney 2023), despite the LLMs' factual errors (Farhi 2023), and

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

in screenwriting, one of the issues raised in the recent WGA writer's strike is the potential replacement of human creative labor with AI (Shah 2023).

OpenAI estimates that 80% of the U.S. workforce will be affected by GPT (Eloundou et al. 2023), requiring significant public policy work. By emphasizing the effect of LLMs on professional human creators, we do not argue that creative careers deserve more protection than others; our focus is on human creators because they are of interest to the CC community and at risk of from our particular research.

To cause these problems, transformer models need not meet traditional CC benchmarks. Their work need not be indistinguishable from human's, or novel and valuable; they need not be autonomous; and they need not take on tasks an "unbiased observer" would deem creative. Transformer models can produce interesting stories when used co-creatively by skilled users (Ippolito et al. 2022), but the AI stories causing problems for Clarkesworld were submitted by scammers and had no artistic value: it is their sheer volume and imperviousness to automated detection that made the magazine's submissions process unworkable (Clarke 2023). To cause economic disruption, generative systems must only produce passable-looking content easily enough to be attractive to scammers, or cheaper for companies than hiring humans.

Content moderation and fabrication

Bender et al. (2021) show that generative models easily produce harmful content, ranging from subtle reflection of social bias to outright abuse, harassment, or hate speech. While OpenAI trained ChatGPT to produce less offensive outputs (Ouyang et al. 2022), it did so using low-paid workers in the Global South to identify harmful content (Perrigo 2023). This human-in-the-loop learning is a common practice for various aspects of large AI models but it can be exploitative, particularly for moderation tasks which expose workers repeatedly to violent and pornographic material. Many workers developed symptoms of PTSD (Perrigo 2023).

Even after training, ChatGPT produces spurious medical advice and other forms of misinformation (Birhane and Raji 2022). This may be an unsolvable problem for transformers, which do not understand the meaning of their output: their high performance on benchmark tests is due to use of statistical cues, not comprehension (Niven and Kao 2019).

Many CC researchers remove offensive content from their system's results by hand. At the scale of GPT-4, this is not possible. Until a radically different method of content moderation is discovered, all researchers working on models of this size face a choice between exploiting workers like this or allowing their AI to generate potentially unlimited hate speech, as in the case of Microsoft's Tay chatbot (Wolf, Miller, and Grodzinsky 2017). The impossibility of ethically moderating content at this scale is, itself, an argument against the use of LLM-sized models.

Public versus private science

Another issue with transformer models is the extent to which research is not peer reviewed. Most OpenAI papers are released on the arXiv. The white paper for GPT4 was released directly by OpenAI, and has no details about GPT-4's data set, training, parameter counts or efficiency (OpenAI 2023).

Reproducibility is thus nearly impossible, as is systematic critique of a model's weaknesses. And companies working on LLMs have worked to stifle such critique. Google famously fired both leads of its ethical AI team after they criticized Google's LLM (Schiffer 2021), and Microsoft cut an ethics group at the exact time it expanded its relationship with OpenAI (Schiffer and Newton 2023).

Researchers at publicly-funded universities struggle to replicate corporate LLM research, both due to the prodigious size of these models and due to ethical concerns. Researchers might be interested in content moderation, for instance, but current content moderation techniques would present difficulties at a university Research Ethics Board. Corporations can build these models and write papers about their outcomes regardless of ethical concerns. While some journals and conference require that their research satisfies ethical standards, the use of arXiv or other non-peer-reviewed venues frees non-academic developers from this constraint.

Academics collaborating with corporations have also avoided scrutiny. One study tested an LLM-based mental health intervention on suicidal teenagers without informed consent. Because the intervention had been designed and implemented by a startup, and the university researchers only analyzed data after the fact, the study was considered to be "non-human subjects research" and the REB did not enforce any protections (Xiang 2023a).

The result is a situation where academic researchers *cannot* reproduce transformer models and *cannot* work at improving their basic mechanisms, but *can* collaborate with the companies who build them, as long as they turn a blind eye to ethical concerns.

A past example of public versus private science

This is not the first time big science has experienced a tension between public and private ownership. The Human Genome Project (Consortium 2001) was an international consortium of researchers, mostly from the US and UK. In 1998, Celera Genomics was founded in part to speed up sequencing. Celera used publicly-generated sequencing data along with its own sequencing to piece together a potentially more accurate human genome, since its input data was a superset of the public data. Users of Celera's data could search for matches to a query, but could not download the full draft sequence or train models on Celera data. Celera's researchers published a paper (Venter et al. 2001), which appeared in the same week as the HGP's (Consortium 2001). For the HGP, all data was publicly available; for Celera, the data was protected by a licensing agreement, and follow-up research was tightly controlled. Fortunately, Celera's advantage over the public-sector project soon eroded. Developers needed sequencing information that Celera did not release, so most researchers analyzed the public data.

Does it matter when scientific data sets are privately held, despite deriving from the work of the world? We argue that it does, in particular for transparency. As people highlight ethical troubles with privately-held LLMs, all of their work is done in the proprietary space of the companies, and the companies need not respond accountably.

Literature review

ICCC is the largest international conference devoted to computational creativity. Did we predict any of the current issues caused by transformers? Did we see the coming storm?

We felt that, overall, ICCC researchers did not predict the current state of affairs. As a test, we conducted a literature review of ICCC papers between 2017-2022, *i.e.*, since the original transformer paper (Vaswani et al. 2017).

Framework of the literature review

We studied papers about text generation or media generation based on text, where transformers have caused the most disruption; papers about ethics and/or the nature of creativity; and general CC reviews. Both authors used Covidence to screen each paper for relevance to these topics. We evaluated each paper on the following questions:

- Did the paper mention neural networks? Did it mention transformers? Was either topics the paper's main focus?
- Did the paper attempt to predict how its area of CC was going to develop in the future?
- Did the paper mention ethics? If so, did it mention any of the specific ethical issues that are the focus of this paper? What other ethical issues were discussed?

Results of the literature review

Figures 1 and 2 show our major findings:

- While neural networks have always been studied in CC, there was a sharp increase in their mention and use in 2020, more recently driven by the rise of transformers. By 2022, most reviewed papers mentioned neural networks, and 43% had neural networks as their central topic.
- Between 25% and 50% of papers studied discussed ethics, but most discussions were brief and concerned other ethical topics than the ones that we screened for, such as how to conceptualize machine ethics or promote social causes. Each of the specific ethical topics we screened for was discussed by a handful of authors at most.
- Exploitation of content moderators was never mentioned.
- We also counted each paper's references taken from arXiv; this ranged from 0 to 54%, with median 0% across all included categories and 2% for text generation papers, but 23% for papers whose primary topic was transformers and 24% for media generation from text prompts.

Discussion

The CC community is not unaware of the *technologically* disruptive potential of transformers; there has been a sharp increase in interest in their use. But this has not been accompanied by a similar increase in attention to their ethical problems.

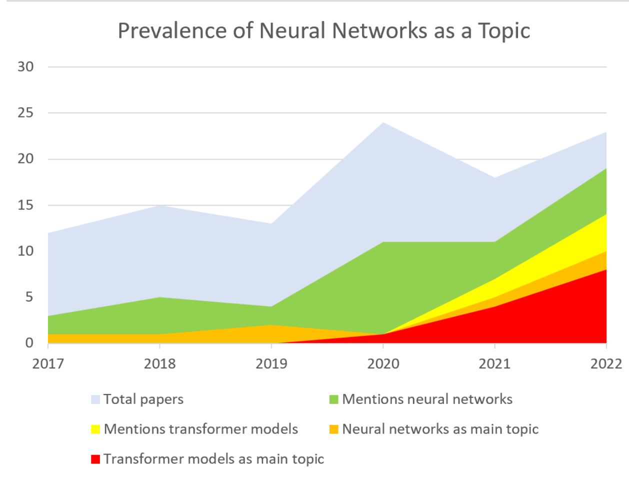


Figure 1: Papers in the screened categories that mentioned neural networks. The areas in this chart are overlapping, *not* stacked. Mentions of neural networks sharply increase beginning in 2020; those of transformer models do so in 2021-22.

This is not to say that attention to the ethical problems was missing entirely. Most of our topics were studied by a few researchers. For example, Bodily and Ventura (2020) discuss consequences for creative humans who feel surpassed by computers. Brown et al (2021) and Gordon et al (2022) analyze copyright issues. Loughran (2022), among others, discusses CC training set bias. Mirowski et al (2022), developing a CLIP-based collage system, incorporate concerns for human autonomy and copyright into their design. But these researchers are a minority. Their recommendations were not been taken up by the broader community, and certainly not at rates that matched the general dramatic increase in use of transformers. Nor did any predict the level of widespread social alarm that we currently see.

It is possible that researchers also raised ethical and social concerns in venues other than ICCC, or in informal discussions. Early signs show that there may be a greater focus on ethics at this year's conference. Nonetheless, the discrepancy between the use of transformers and the attention paid to ethics, in the papers that ICCC published before the explosion of public interest in this topic, is striking.

Our count of arXiv references is not a stand-in for paper quality; many papers full or arXiv references are thoughtful and inventive. (Indeed, this manuscript cites arXiv, news sources, and blogs!) However, the differences in this metric across categories suggest that in certain areas, the state of the art forces researchers to rely on non-peer-reviewed claims.

In light of the social disruption caused by transformers, some of the CC community's usual foci feel less urgent. Issues such as autonomy and embodiment are orthogonal to social impact; transformers cause these impacts regardless. Many impacts have nothing to do with the models' inner workings and everything to do with their ease of use at scale.

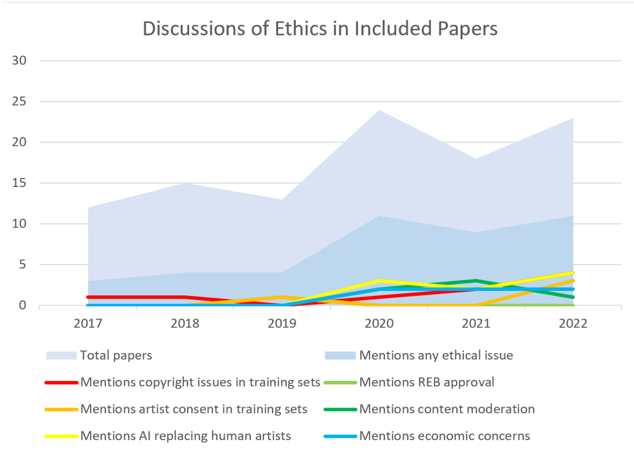


Figure 2: Number of papers in the screened categories that mentioned ethics. The areas and lines are overlapping, *not* stacked. It is a low, steady rate over time.

Conclusions and Recommendations

Transformers are the most quickly developing area in CC today. Yet, as we have seen, they produce harmful misinformation, exploit content moderators, harm the environment and are produced by opaque corporations that silence criticism. They make questionable use of copyright exceptions to gather training data for purposes that economically affect the artists on whose work they train, and they raise fears of human artists being replaced altogether. The hype and the alarm are overwhelming. What is a CC researcher to do?

One option is to study transformers ourselves, but if we do this, we must do so critically, with intense attention to their social and economic effects; we cannot become shills. ICCC is devoted to all aspects of computational creativity; social impacts are one such aspect, and they are exploding. To avail ourselves of the benefits and interesting uses of transformers, without proper and significant attention to these impacts, is a woeful imbalance.

As academics we lack direct power over corporations and governments, but we have a respected voice. We can rally against the excesses of corporate AI, point out its drawbacks, and suggest mitigations or alternatives. As ”creative AI” becomes a public policy issue, more of us must focus on these roles.

Another option is to avoid transformers altogether. There are arguments in favor of this option; Bender et al (2021) discuss the “opportunity cost” of pouring scientific, financial, and material resources into transformers instead of using them to develop better alternatives. However, this option is not the easy out that it may appear. We must be aware that transformers are now the public face of creative AI - the first and sometimes only thing that a member of the public thinks of when they think of what we do. In this environment, if we develop a generative system that is not a transformer, it is up to us to clearly differentiate it from a transformer. We should think about how our own models can avoid the ethical pitfalls into which transformers have already fallen, and how we can make this difference clear to a frightened or skeptical audience.

At the least, we must be aware of the social effects of our research. Beyond writing about ethics in theory, we must incorporate ethics into our process, for example by adopting the recommendations of Bender et al (2021): thoroughly document training datasets, identify stakeholders at risk, and re-align research goals around a system’s sociotechnical role.

We have an advantage in our emphasis on Process and Press, not merely Product (Jordanous 2016). We should take care not to lose this advantage. The social impact of CC systems has reached a crisis point, and is the most urgent issue in CC today; we should treat it accordingly.

Acknowledgments

The work of D.G.B. is supported by an NSERC Discovery Grant. We appreciate helpful conversations with Max Peeperkorn, Lauren Byl, and Maura Grossman as well as the clarifying comments of our anonymous peer reviewers.

Author contributions

Author C.E.L. designed the literature review, outlined the paper, and drafted most of it. Author D.G.B. drafted the two sections on public vs private science. Both authors collaborated closely on ideating the paper, performing the literature review, and revising the text.

References

Belanger, A. 2023. Getty sues Stability AI for copying 12m photos and imitating famous watermark. *Ars Technica*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proc. 2021 ACM Conf. on fairness, accountability, and transparency*, 610–623.

Bensinger, G. 2023. ChatGPT launches boom in AI-written e-books on Amazon. *Reuters*. Feb 21.

Birhane, A., and Raji, D. 2022. ChatGPT, Galactica, and the progress trap. *WIRED*. Dec 9.

Bodily, P. M., and Ventura, D. 2020. What happens when a computer joins the group? In *Proc. ICCC’20*, 41–48.

Brown, D.; Byl, L.; and Grossman, M. R. 2021. Are machine learning corpora “fair dealing” under Canadian law? In *Proc. ICCC’21*.

Clarke, N. 2023. Editor’s desk: Written by a human. *Clarkesworld*. April.

Consortium, I. H. G. S. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

Cropley, D. H.; Kaufman, J. C.; and Cropley, A. J. 2008. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal* 20(2):105–115.

Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Farhi, P. 2023. A news site used AI to write articles. it was a journalistic disaster. *The Washington Post*. Jan 17.

Gordon, S.; Mahari, R.; Mishra, M.; and Epstein, Z. 2022. Co-creation and ownership for AI radio. In *Proc. ICCC'22*.

Ippolito, D.; Yuan, A.; Coenen, A.; and Burnam, S. 2022. Creative writing with an AI-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*.

Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194–216.

Lemley, M. A., and Casey, B. 2020. Fair learning. *Texas Law Review* 99:743.

Loughran, R. 2022. Bias and creativity. In *Proc. ICCC'22*.

Mirowski, P.; Banarse, D.; Malinowski, M.; Osindero, S.; and Fernando, C. 2022. Clip-clop: Clip-guided collage and photomontage. In *Proc. ICCC'22*.

Niven, T., and Kao, H.-Y. 2019. Probing neural network comprehension of natural language arguments. In *Proc. ACL*, 4658–4664.

OpenAI. 2023. GPT-4. Mar 14.

Ouyang, L.; Wu, J.; Jiang, X.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS* 35:27730–27744.

Perrigo, B. 2023. Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *TIME*. Jan 18.

Scalzi, J. 2023. OMG is the AI coming for my job?!???!??!!!?!?!? *Whatever*. Feb 23.

Schaub, M. 2022. Cover of Paolini book may contain AI-created image. *Kirkus Reviews*. Dec 16.

Schiffer, Z., and Newton, C. 2023. Microsoft lays off team that taught employees how to make AI tools responsibly. *The Verge*. Mar 13.

Schiffer, Z. 2021. Google fires second AI ethics researcher following internal investigation. *The Verge*. Feb 19.

Shah, S. 2023. The writers srike is taking a stand on AI. *Time*. May 4.

Sobel, B. L. 2017. Artificial intelligence's fair use crisis. *Columbia Journal of Law & the Arts* 41:45.

Sweney, M. 2023. Mirror and Express owner publishes first articles written using AI. *The Guardian*. Mar 7.

Vaswani, A.; Shazeer, N.; Parmar, N.; et al. 2017. Attention is all you need. *NeurIPS* 30.

Venter, J. C.; Adams, M. D.; Myers, E. W.; et al. 2001. The sequence of the human genome. *Science* 291:1304–1351.

Wolf, M. J.; Miller, K.; and Grodzinsky, F. S. 2017. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *ACM SIGCAS* 47(3):54–64.

Xiang, C. 2023a. 'Horribly Unethical': Startup experimented on suicidal teens on social media with chatbot. *Vice*. Mar 7.

Xiang, C. 2023b. Legendary sci-fi magazine halts submissions amid deluge of AI-written stories. *Vice*. Feb 21.

Zirpoli, C. T. 2023. Generative artificial intelligence and copyright law. *Congressional Research Service*.

Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation

Ken Déguernel[1,2] and **Bob L. T. Sturm**[1]
[1]Royal Institute of Technology KTH, Lindstedtsvägen 24 SE-100 44 Stockholm, Sweden
[2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Abstract

This paper surveys 27 published studies exploring bias in the evaluation of computational creativity. These studies look specifically at the involvement of AI, in generating music, images and graphics, poetry, and news articles. While some have found evidence of bias (43%), others find no bias (27%), or show modulation of bias through socio-cultural factors (30%) resulting in a lack of consensus on this issue. We argue for the importance of taking into account socio-cultural context when considering such biases in these creative pursuits. What styles do the artefacts belong to? Who are the participants involved in the study, and what are their relationships to the styles at hand? We discuss the implications of such considerations for future research in computational creativity. We propose some safeguards when conducting a study on bias in the evaluation of computational creativity, and propose directions to study more specifically when, and with whom it can be observed.

Introduction

Artificial intelligence (AI) has been applied in a wide variety of artistic fields such as poetry (Hämäläinen and Alnajjar, 2019), painting (Ramesh et al., 2021), comedy (Strapparava and Stock, 2011), and music (Herremans, Chuan, and Chew, 2018), and other more technical fields such as journalism (Broussard et al., 2019) or programming (Li et al., 2022). Applications of "computational creativity" have reached the attention of the general public through popular tools for generating free-form text (Brown et al., 2020), and generating images from textual descriptions (Rombach et al., 2021).

Human appreciation of creativity and its results is influenced by many factors, such as age, gender, personality, and expertise, but is also influenced by external factors regarding knowledge and context of production (Davies, 2003; Steinbeis and Koelsch, 2009) and socio-cultural factors, such as values and practice. Since knowledge of the production process is an important evaluative criterion (Lamb, Brown, and Clarke, 2018), a bias could exist when it comes to knowing or thinking that an artefact arises from computational creativity. Knowledge about such bias is moreover important when it comes to the evaluation of such creative systems as the appreciation of the artefacts they produce is sensitive to many subjective criteria. While the evaluation of computationally creative systems based on how close the artefacts they produce come to human-created ones can provide valuable insights, it encourages "superficial imitation" (Pease and Colton, 2011), and fails to take into consideration external factors that could trigger some bias in favour or against AI that could influence those results. As such, it is difficult to say whether an artefact generated (even partially) via computational creativity can be evaluated the same way as any other human-created artefact (Ariza, 2009).

One can see the bias discussed above in the frame of *algorithmic aversion*, a phenomenon where individuals have a negative attitude or mistrust towards AI systems (Dietvorst, Simmons, and Massey, 2015). This can manifest in various ways, such as resistance to using tools or services involving AI, scepticism about AI-generated decisions, and concerns about the impact of AI on society (Flick and Worrall, 2022). With a survey of 80 studies, Mahmud et al. (2022) identifies factors linked to algorithmic aversion: algorithmic factors such as the explainability of the algorithm, its presentation and accuracy; individual factors such as personality, psychological factors and familiarity with algorithms; high-level factors such as by whom algorithms are being used (e.g., banks, for-profit organizations) and social influences; and finally, task factors as in what the algorithms are used for. The opposite of this phenomenon is called *algorithmic appreciation* (Logg, Minson, and Moore, 2019).

In this paper, we review 27 papers describing studies in which quantitative analysis is applied to detect and measure bias for the task of music generation, image and graphic generation, and text generation. We propose potential explanations for when bias is (or is not) observed, such as the lack of accounting for contextual factors through the selection of study participants, or the presentation of artefacts with respect to their use. We discuss the implications of these results for future studies on bias against computational creativity, and on the evaluation of such systems.

A survey of contradicting results

This section surveys the results of all studies (to the best of our knowledge) explicitely attempting to measure bias in the evaluation of computational creativity for music generation, graphics and images, poetry, and journalism, in order to observe a variety of media, and both artistic and factual scopes. Table 1 summarizes these 27 publications.

Task	Paper	Style/Topic	N	Reported conclusion
Music	Dahlig and Schaffrath (1998)	German Folk songs	432	Varied
	Moffat and Kelly (2006)	Contemporary, free jazz, Bach	20	No bias
	Friedman and Taylor (2014)	Classical	58	No bias
	Pasquier et al. (2016)	Contemporary	122	No bias
	Jago (2019)	Song	200	Bias
	Hong, Peng, and Willians (2021)	EDM, classical	299	Bias
	Moura and Maw (2021)	Pop, classical	86	No bias
	Aljanaki (2022)	Classical	20	Bias
	Déguernel, Sturm, and Maruri-Aguilar (2022)	Irish traditional music	46	Bias
	Shank et al. (2022)	Classical	136	Bias
	Hong et al. (2022)	Rock, EDM, classical, country	222	No bias
Graphics and Images	Kirk et al. (2009)	Modern art	14	Bias
	Norton, Heath, and Ventura (2015)	Abstract	284	Bias
	Chamberlain et al. (2018)	Abstract, representational	65	Bias
	”	Portrait	349	Varied
	Hong (2018)	Abstract	28	Varied
	Hong and Curran (2019)	Abstract/Psychedelic	288	No bias
	Jago (2019)	Painting	201	Bias
	Ragot, Martin, and Cojean (2020)	Landscape/Portrait	565	Bias
	Wu et al. (2020)	Modern	544	Varied
	Gangadharbatla (2021)	Abstract, representational	530	Varied
Poetry	Wu et al. (2020)	Modern	544	Varied
	Hitsuwari et al. (2023)	Haiku	385	Varied
Journalism	Clerwall (2014)	Sport	46	No Bias
	van der Kaa and Krahmer (2014)	Sport, finance	252	Varied
	Graefe et al. (2018)	Sport, finance	986	Bias
	Waddell (2018)	Political news	311	Bias
	Liu and Wei (2019)	Spot news, interpretive news	355	Varied
	Longoni et al. (2022)	Headlines	3029	Bias
	Lermann Henestrosa, Greving, and Kimmerle (2023)	Popular science	469	No bias

Table 1: Summary of studies on bias in the evaluation of Computational Creativity with their respective topics studied, number of participants (N), and conclusions. “Varied” indicates that the existence of bias is modulated by socio-cultural factors.

Music Generation

One of the first experiments in this area is that of Dahlig and Schaffrath (1993, 1998). A participant listens to a melody and rates the degree to which it is “original” or “computer made” and whether they like it. As stimuli they use computer syntheses of eleven melodies: two German folk melodies and eleven melodies created by mixing up phrases of German folk melodies à la *Musikalisches Würfelspiel*. They report having 432 respondents, drawn from a variety of populations, including musicologists and young students from schools in Germany and China. From the results they conclude, “the biggest number of positive aesthetic evaluations was accorded to melodies regarded to be authentic. Contrariwise, melodies ‘suspected’ to be computer-made got the biggest number of negative evaluations”.

Moffat and Kelly (2006) describes a two-stage listening experiment where participants assess six 1-minute excerpts of music. In the first stage, a participant listens to each excerpt and answers questions such as, “How much do you like this sample?” and “Do you think it was composed by a human or by a computer?” In the second stage, the participant is told the true origin of each excerpt (name of composer or computer system), and is asked questions such as, “Would you buy this piece of music?” Three of the excerpts are of computer-composed music in the styles of “Bach”, “free-form jazz”, and “pieces for strings”, and three are of human-composed music in the same styles. They report from data collected from 20 participants that while they find “there is a common bias against computer-generated pieces”, and that “[i]n almost every case, a piece of music is preferred when it is thought to be human-composed”, they do not observe any significant differences between the rating of liking (stage 1) and enjoyment (stage 2) after a listener is told the origin of the music.

Pasquier et al. (2016) extends the study of Moffat and Kelly (2006), and builds upon past work in evaluating creative systems (Eigenfeldt, Burnett, and Pasquier, 2012). In the study, a participant listens to a music excerpt and rates their perception of it on four dimensions: “Good–Bad”, “Like–Dislike”, “Emotional–Unemotional”, and “Natural–Artificial”. Each participant listens to and rates each excerpt twice in the experiment, but in one of three different conditions. A participant in the “fully informed condition” is told about the origin of each excerpt. A participant in the “fully naïve condition” is never told about the origin of each excerpt. And a participant in the “revealed condition” is only told about the origin of each except after listening to and rating all excerpts once. They use 1-minute excerpts of six “contemporary string quartets”, three composed by a human and three generated by an AI system. They report from 122 participants (university students) that “[w]hile our results do indicate a negative effect of the knowledge of computer authorship on listener judgements, this effect is not significant”.

Friedman and Taylor (2014) describes a study where a participant listens to a music recording, and then rates several qualities, e.g., arousal, liking and quality. The participant then decides whether the piece was composed by computer or human, and whether it was played by computer or human. Each participant is assigned to one of two conditions: either the participant is explicitly told every music recording was composed and performed by a computer; or the participant is explicitly told every music recording was composed and performed by a human. The study uses synthesized recordings of four human-composed classical piano pieces of between 1.6 and 3.4 minutes duration. From an analysis of over 190 participants (undergraduate psychology students), they conclude, “the perception that the music was computer-generated did not significantly alter participants' emotional responses or their judgments of the quality of what they had heard."

Jago (2019) presents a study where a participant listens to a 30-second recorded music excerpt and rates their perception of the "authenticity" of the work. In one condition, the participant is told the work is by a particular person. In another condition, the participant is told the work is by a particular AI. Four different music excerpts are used, each generated by the same AI system; but a participant only rates one excerpt. Based on the responses of 200 participants (from Amazon Mechanical Turk (Mturk) users in the USA), Jago (2019) concludes that the participants "believed that human work was more authentic, compared to an artificially intelligent algorithm's otherwise-identical work."

Moura and Maw (2021) describes a study where a participant reads a narrative about the music they will hear, then listens to two 1.5-minute music recording excerpts, and then answers questions about the experience. All participants listen to the same music, but each is assigned to one of two groups, corresponding to a particular narrative. One narrative states the music is composed by AI, and the other describes emotions and experiences reflected in the music, implying a human composed the music. The two excerpts are "pop-rock" and "classical" styles, each arising from human-AI collaboration. Based on the responses of 86 participants (German university students) they report no significant differences in responses between the two groups for either music excerpt, and conclude, "listeners' awareness of the nature of the composition process (human versus AI) posed no significant impact on participants' perceptions towards the songs [...] regardless of the different music genres."

Aljanaki (2022) discusses a study where a participant listens to recordings of two pieces for piano and rates each. All participants listen to the same music, but each participant is assigned to one of two groups: in one the real origin of each piece is given; in the other the origin is reversed. One piece is modern and composed by a human, and the other is composed by a machine, "reminiscent of romantic period in classical music". From the responses of 20 participants ("non-musicians"), Aljanaki (2022) concludes that the difference between responses of the groups was not significant.

Déguernel, Sturm, and Maruri-Aguilar (2022) describes a study where a participant listens to six music recordings, rates their liking of each, and then listens to them again rating their belief of each being composed by a computer. The six music recordings feature the same musician playing six different computer-generated "double jigs", (a form of Irish traditional dance music). Based on the responses of 46 participants (practitioners of Irish traditional music), they conclude that the practitioners "tend to like more the tunes they deem hardly likely to be composed by an AI. Alternatively, the more they report liking a tune the less they report believing the tune is AI-composed."

Hong, Peng, and Willians (2021) describes an experiment where a participant listens to a music recording and is told it is composed by either AI or a human, and is then asked to evaluate the music. Each participant is given only one of the four pieces and one of the possible origins. Four AI-composed pieces are used, two of the type "classical" and two of the type "EDM" (electronic dance music), each generated by the same AI system. Based on the responses of 299 participants (found using Mturk), they conclude, "accepting the creativity of AI is a prerequisite for a positive evaluation of its artistic merit ... [A]n unwillingness to accept AI products blocks appreciation."

Hong et al. (2022) presents a study where a participant reads a "mock" news article about an AI music generation system, then listens to a piece of music presented as composed by that system, and finally rates their experience of the piece. There are four different news articles and four different pieces. Each news article describes a different level of anthropomorphism and algorithmic independence of the AI music generation system. The four different pieces are composed by the same AI system, but in the styles "rock", "EDM", "classical" and "country." Each participant is randomly assigned a news article and a piece. Based on the responses of 222 participants (found using Mturk), they conclude that neither aspect have any significant impact on the ratings of the music.

Shank et al. (2022) presents three studies investigating the relationship between reported music liking and belief in AI authorship. In the first study a participant listens to twenty 15-second excerpts of human-composed music, and after each is asked whether it was composed by a human or AI and their confidence so, and is finally asked to rate how much they like the music. Each participant is given excerpts of either the music type "classical" or "electronic". Based on the responses of 295 participants (found using Prolific), they conclude that "music that was perceived as being composed by an AI was liked less than music that was perceived as being composed by a human". In a second study, a participant listens to eight 15-second excerpts of human-composed music, and after each is asked to rate their liking of it and its qualities. These specific excerpts were selected based on the responses to the previous study: four electronic music excerpts were selected as sounding the most "AI", and four electronic music excerpts were selected as sounding the most "human". The participant is assigned to one of three conditions. In the first, they are told all excerpts are composed by AI composing software. In the second, they are told all excerpts are composed by various composers. In the third, they are not told of the origin of the music. Based on the responses of 399 participants (found using Prolific), they do not find a significant effect of the purported origin on participant liking. They then present a third study where a participant listens to eight 15-second excerpts of human-composed music, and after each is asked to rate their liking of it and its qualities. These specific excerpts were selected based on the responses to the first study: the classical music excerpts sounding the most "human." The participant is told beforehand that some of the excerpts were composed by AI software. Each excerpt is presented as being composed by a specific person or a specific AI system. Based on the responses of 136 participants (found using Prolific), they conclude that "participants rated the music as both lower quality and liked it less if it were purportedly composed by an AI."

The conclusions from these 12 papers show a clear lack of consensus on the existence of a bias in the evaluation of music generation systems. This could be explained by the use of different musical style (although different results have been found for classical music (Friedman and Taylor, 2014; Shank et al., 2022)), and the use of different criteria of evaluation and presentation of the algorithm (Hong, Peng, and Willians, 2021). Déguernel, Sturm, and Maruri-Aguilar (2022) also suggests a potential role of expertise and familiarity as a modulating factor of such a bias.

Graphics and images

Kirk et al. (2009) presents a study where a participant views a digital image of an abstract painting together with a text label showing its origin, and rates its pleasantness (aesthetic rating scale). Each participant is told they will see 200 abstract paintings, that half of them are from a famous gallery, and that half are generated by the experimenter using computer software. The 200 digital images were "selected from online sources" by the experimenters, and all appear to be human-created. Based on the responses of 14 participants (university students in Denmark), they conclude that "images under the gallery label were rated as having a significantly higher mean aesthetic value than those carrying the computer label."

Norton, Heath, and Ventura (2015) discusses a study where a participant views a pair of images (processed digital photographs) – one labeled as created by a human and the other labeled as created by a computer program – and selects the one they believe is a better image. All images for fifteen pairs were generated by the same computer program, and selected by the experimenters. From 330 responses collected online, they conclude that there was "a small but substantial bias either towards humans or against [the algorithm]."

Chamberlain et al. (2018) describes a study where a participant is shown in random order sixty digital images and is asked after each how much they like it, and then shown the images a second time and asked after each if they believe the image is man-made or computer-generated. A participant in a reversed condition is asked first if they believe an image is man-made or computer-generated, and then how much they like it. Half of the images were selected by the experimenters from online "computer art databases" being of types "abstract" or "representational", and the other half were of man-made artwork of the same types. Based on the responses of 65 participants (students and staff at KU Leuven), they conclude that for either condition "images that were categorized as computer-generated were rated as visually less pleasing." Chamberlain et al. (2018) describes a second experiment where a participant evaluates drawn human portraits made by a robot artist (a table-mounted animatronic arm holding a pen which makes marks on a piece of paper). Some participants see the robot and its artworks; some participants are just told about the robot and shown the artworks; and some participants are only shown the artworks and not told anything about them. Each participant answers a survey about their aesthetic responses. Based on the responses of 349 participants in the three conditions (attendees of the art gallery, and KU Leuven students and staff), they conclude the bias observed in the first experiment "can be moderated by interaction with the agents of the artwork. The presence of the robotic artists had a strong positive impact on aesthetic evaluations of the resulting artworks."

Hong (2018) describes a focus group in which participants view a digital image of an artwork and discuss questions about art and the involvement and relevance of AI. In one condition the group is told that the image they are viewing was produced by AI. In the other condition, the group is told it was produced by a human. Both groups view the same image, which was created by a human artist. From the discussion of the 14 people in each group (students at the University of Southern California), Hong (2018) concludes that the group being told the image they are viewing was produced by AI had "a stronger tendency toward the belief that AI cannot produce art," and that "one way to diminish a negative stereotype toward artificial intelligence being creative is to successfully persuade the public its autonomy" – which echoes the finding with the perception of robot artists in Chamberlain et al. (2018).

Hong and Curran (2019) presents a study where a participant views a digital image of an abstract artwork and then rates it along eight dimensions, e.g., originality, composition, and aesthetic value. There are four groups of participants, crossing factors of attribution knowledge (being told the images are created by AI, or not being told anything about human or AI authorship), and image source (images are generated by AI, or images are human created). Participants in the groups being told images are created by AI view the same set of six images; and the participants in the other groups view a different set. Six of the twelve images used are generated using three AI systems. The remaining images are of six human-made paintings, selected by the authors for sharing stylistic and thematic similarities with the AI-generated images. In each set of six images viewed by a group, half are from AI systems. From the responses of 288 participants (from Mturk) they conclude that "[the] evaluation of aesthetic value is done independently from bias related to the artwork and its artist."

Jago (2019) study, presented in the previous section, also have a participant sees and rate a digital image of a painting with the same procedure described above. Based on the responses of 200 participants (from Mturk users in the USA) Jago (2019) concludes again that "they believed that human work was more authentic, compared to an artificially intelligent algorithm's otherwise-identical work."

Wu et al. (2020) presents a study exploring the explicit and implicit attitude towards AI-generated paintings. Participants are shown a digital image of either a human- or AI-created painting, then asked to rate it on quality, imaginativeness, spatial presence, empathy, competence, and finally to rate their attitude towards AI. To take into account the implicit bias, participants are given an alleged human or AI origin for the piece they are evaluating. Based on the responses from 251 U.S. participants and 293 Chinese participants they report that U.S. participants were more critical to AI-generated art compared to human-generated content both explicitly and implicitly, whereas Chinese participants exhibited overtly positive attitudes towards AI-generated content, yet their implicit acceptance of it was lower than that of human-generated content.

Ragot, Martin, and Cojean (2020) discusses a study in which a participant views a digital image of an artwork and rates it along four dimensions, e.g., liking and novelty. Participants in one group are "primed" with information that the artworks they will see were created by "some artificial intelligence", and in the other group that the artworks were created by "some artists". Each participant views 8 images, selected at random by the authors from 40 curated images: "10 portraits by AI, 10 landscapes by AI, 10 portraits by humans, and 10 landscapes by humans". Both human and AI artists were involved, Based on responses of 486 participants (from Mturk) they conclude "the artworks presented as AI-generated paintings were significantly less liked and were perceived as less beautiful, novel, and meaningful than paintings presented as drawn by a human."

Gangadharbatla (2021) describes a study where a participant views a digital image of an artwork and then rates nine characteristics of it, including creativity, aesthetic value and financial value. In one condition, a participant is given prior information that the images were generated by AI without human involvement. In another condition, the prior information relates to the human production of the artworks they will see. Each participant views the same four images of two types of art: "representational" and "abstract". One work in each type is human-created and the other is AI-generated. Based on responses of 530 participants (from Mturk) they conclude that "attribution knowledge [plays] a significant role in influencing individuals' evaluations of artwork."

The conclusions from these 9 papers similarly display a lack of consensus on the existence of a bias in the evaluation of artwork generation systems, with different results for the same types of artworks. Several factors of modulation of bias are found in those studies however. Personal factors such as culture, identified by Wu et al. (2020), algorithmic factors depending on how the system is presented or observed, identified by Hong (2018); Chamberlain et al. (2018), and contextual factors such as where the experiment is conducted, identified by Chamberlain et al. (2018).

Poetry

Wu et al. (2020) presents a study exploring the explicit and implicit attitude towards AI-generated poems, using the same procedure as for the graphics generation described above. Based on responses from 251 U.S. participants and 293 Chinese participants they conclude the same: that U.S. participants were more explicitly and implicitly critical to AI-generated poetry compared to human-generated content; and Chinese participants exhibited overtly positive attitudes towards AI-generated poetry, yet their implicit acceptance of it was lower than that of human-authored poetry.

Hitsuwari et al. (2023) describes a study consisting of two blocks: first, a rating block where a participant rates their liking of haikus according to 21 criteria such as beauty, valence, arousal, and novelty; and then a discriminating block where a participant is asked whether they think the haiku was created by AI or a human, and what criteria they use to make their decision. In one condition, a participant rates poems first and then predicts the author. In the other condition, these tasks are reversed. Stimuli are either human-made, AI-made, or made with a "Human in the loop". Based on the responses from 385 participants (Japanese recruited through CrowdWorks), they report that "task order (i.e., prior knowledge about whether the work was produced by AI) did not affect the evaluation of the beauty of haiku". However, the more beautiful a haiku was rated, the more likely it was believed to be created by a human.

Both those studies show a modulation of bias in the evaluation on poetry generation systems. On the one hand, Wu et al. (2020) identify culture as a personal factor and on the other hand, Hitsuwari et al. (2023) identify the presentation of the systems as an algorithmic factor modulating bias in evaluation.

Journalism

Clerwall (2014) describes a study where a participant reads a written account of a sports game, evaluates the article according to 12 descriptors (e.g., objective, trustworthy, and informative), and then is asked whether they think the text is human- or computer-written. Each account is either generated by a computer or written by a journalist. Based on the responses from 46 participants (undergraduate media students), Clerwall (2014) reports no significant differences on how the groups evaluated or perceived the articles.

van der Kaa and Krahmer (2014) replicates the study of Clerwall (2014) with news topics of sports and finance. In their study, however, participants are given an alleged source for the article: either a journalist or a computer. Participants rate the article according to the same 12 descriptors. Based on the responses from 188 Dutch news consumers and 64 professional journalists, they conclude there were "no differences in the perceptions of news consumers" depending on authorship attribution, and that "news consumers have no strong negative or positive feelings toward computer-written news". On the other hand, "[j]ournalists perceive themselves as more trustworthy compared to their 'computer colleagues"', showing an impact of expertise. They also note a difference in the perceived level of trustworthiness depending of the topic of the article.

Graefe et al. (2018) replicates the study of van der Kaa and Krahmer (2014). Based on the responses of 986 German-speaking participants (recruited through SoSci Panel), they report that "articles are consistently perceived more favorably if they are declared as written by a human journalist, regardless of the actual source".

Waddell (2018) discusses two studies in which participants rate the accuracy and credibility of news article. In the first study, participants read a data-driven news article about politics attributed to a known news source. Participants are randomly assigned a condition in which the article is attributed to a specific journalist or to a "robot reporter". The second study replicates the first one but participants read articles about the weather, stock market, and science, and are also asked to fill in a "robot recall" questionnaire in which they are asked to recall a film or a show which involve a robot as a main or supporting character, and answer questions about how "good or bad", "human-like", this character is. Based on the responses from 129 in the first study and

182 participants in the second study (all recruited through Mturk), Waddell (2018) reports that "news attributed to a machine is perceived as less credible than news attributed to a human journalist". This effect is still present after the "robot recall", although it is slightly modulated by it.

Liu and Wei (2019) describes a study with two news organisations two news types (sport and interpretive news), and two alleged writers (AI or human). Participants first indicate their political values according to a questionnaire, and then are asked to read one randomly selected news article using the templates from the newspapers' websites and with the alleged identity of the writer indicated at three places in the article. Participants then rate their emotional involvement and perception of the news article. Based on the responses from 355 U.S. participants (recruited through Mturk), they report that AI author attribution induces less emotional involvement, is perceived of less expertise, but is also perceived as more objective. Moreover, "for a media organization whose news was more trusted, utilizing news-writing bots enhanced perceived news objectivity. Otherwise, employing bots further reduced perception of the writer's trustworthiness and expertise", showing an effect of the context and of the participants' political opinions.

Longoni et al. (2022) describes two studies in which participants rate the trustworthiness and accuracy of news headlines. In the first, participants are randomly assigned to a condition where they see the news items tagged as written by AI, or a condition where they see the news items tagged as written by human. In the second study, participants see news items tagged both ways. Some headlines are true and some are false. Based on the responses from 3029 participants in the first study, and 1005 participants in the second study (all recruited through Lucid), they report that the effect of source attribution has a significant effect on trust and on perceived accuracy.

Lermann Henestrosa, Greving, and Kimmerle (2023) presents a study testing if the credibility and trustworthiness of popular science articles can be influenced by human or AI attribution of authorship. Participants are given a popular science article with either a neutral or evaluatively positive tone, and are asked to rate it on 19 criteria judging message credibility and 12 criteria judging perceived trustworthiness. In one condition, participants are told the article was written by a journalist; in the other, they are told the article was written by a computer algorithm. In both cases, alleged sources are provided and participants are told that the articles had been published in a reputable newspaper. Based on the responses from 469 participants (recruited using Prolific), they report that although the tone of the article had a significant impact on the way articles were perceived, this effect is independent of authorship attribution, which has no significant impact on the evaluation of credibility and trustworthiness. They note, "Although participants made a clear difference in how they perceived the alleged authors, this difference was not at all reflected in their evaluation of the message"

The conclusions from these 7 papers show a lack of consensus on the existence of a bias in the evaluation of news article generation systems. Several factors of modulation of bias are found: Contextual factor such as the topic of the article, as found by van der Kaa and Krahmer (2014), and personal factors such as expertise, identified by van der Kaa and Krahmer (2014), and political stance, identified by Liu and Wei (2019).

Discussion: The importance of the socio-cultural context

The previous section refers to 27 papers related to measuring bias in the evaluation of Computational Creativity. It is apparent that there is no clear consensus. Although it is possible that the differences in the outcomes is only due to differences in the methodologies, we believe that these discrepancies are better explained by the different socio-cultural contexts in which the studies were conducted. What styles/topics do the evaluated artefacts belong to? Who are the participants involved in the study? And what are their relation to the styles/topics at hand? In this section, we describe how socio-cultural factors influence art appreciation and our relation to creativity, and how this can be a dominant factor in the evaluation of Computational Creativity.

Contextual factors

Art appreciation is influenced by many properties of the artefact itself (Koelsch, Vuust, and Friston, 2019; Obermeier et al., 2013; Hagtvedt, Patrick, and Hagtvedt, 2008), and can be modulated by personal individual factors (Orr and Ohlsson, 2005; Dubnov, Burns, and Kiyoki, 2016; Hitsuwari and Nomura, 2022). There is empirical proof that knowledge of extra-artistic factors influences the way we perceive art (Leder and Nadal, 2014; Greasley and Lamont, 2016). For instance, Brieber et al. (2014) shows that the setting in which art is experienced influences one's appreciation of it. Art is found more interesting and viewed longer in a museum than in a laboratory setting. Similarly, North, Hargreaves, and Hargreaves (2004) observes that people change their listening habits depending on the time, the activity they are doing, or their location. Flôres and Ginsburgh (1996) shows that the order of music performances had a significant correlation with the ranking of the professional juries in a competition, given performances occurring at the end of the competition an advantage. These kinds of contextual factors also have an impact even when it is only based on belief. For instance, Lauring et al. (2016) shows that for 'art-naïve' students, social priming, i.e., saying that a group of other students or art professionals rated positively or negatively an artwork, or giving alleged price information about an artwork, has a significant impact in their liking rating. Similarly, belief that a piece of music is composed by a well-established artist (Fischinger, Kaufmann, and Scholtz, 2018) or performed by a renowned musician (Kroger and Margulis, 2016) bias a listener's reported appreciation.

Culture and expertise

As described by Lubart (2010), the definition and conceptual boundaries of creativity is dependent on culture, which define on the one hand what and who can be considered creative, but also the "why and how" of creativity: "Culture is omnipresent, and for this very reason its impact is often underestimated." The impact of this have been raised recently in the scope of AI ethics by Huang, Sturm, and Holzapfel (2021) regarding applications of AI to music, showing once again the importance of culture in the "why and how" of Computational Creativity.

Cultural familiarity has been shown to have an impact on the perception and appreciation of specific characteristics of art. For instance, Maher (1976) shows that musical intervals that would be considered very dissonant in Western culture appear in Indian classical music, and trigger different responses depending on the familiarity of the participants. A more extreme case of this phenomenon has been shown by McDermott et al. (2016), who observe that "consonant" or "dissonant" harmony is not a characteristic that matters for the music appreciation of native Amazonians. Lahdelma and Eerola (2020) recommend controlling for cultural familiarity and musical expertise for studies involving the perception of music dissonance.

Expertise is another factor that influences art perception and appreciation. Winston and Cupchik (1992) studies the aesthetic assessments of art-naïve and experienced students when shown "popular art" and "high-art paintings". They show that while art-naïve students prefer popular art, and experienced students prefer high-art paintings, the evaluation criteria used by each group are different: art-naïve students report more their emotional responses to artworks, while experienced students focus more on objective and structural properties of the artworks. Pearce (2015) describes a similar phenomenon for music, showing that although musical expertise is not significantly correlated to emotional experience, it has an impact in the processing of long-term musical structure, and on the aesthetic judgement of consonance and dissonance, and of musical complexity.

Darda and Cross (2022) studies the impact of cultural familiarity on the evaluation of Indian and Western visual art (painting and dance). Indian participants (21 experts, 24 non-experts) and Western participants (21 experts, 26 non-experts) are shown abstract and representational paintings and dance videos belonging either to Indian culture or Western culture. Participants are asked to rate the stimuli according to familiarity, complexity, evocativeness, abstractness, technical competency, beauty and liking. They report that cultural familiarity creates a in-group bias for dance (although, the same is not found for painting) and that there is a preference for representational art. However, the in-group bias is modulated by expertise as it is only found in art-naïve participants. Similarly, the preference for representational art is also modulated by expertise but only for Western participants. This study shows the intertwined relationship and disparity between cultural familiarity and expertise, which creates a complex system to consider when evaluating bias.

In regards to the evaluation of Computational Creativity, our survey shows that cultural familiarity and expertise has an impact. Wu et al. (2020) shows the impact of cultural background on the explicit and implicit attitude towards AI-generated poems and paintings, reporting a difference between U.S. and Chinese participants regarding their explicit acceptance of AI-generated contents and general attitude towards it. van der Kaa and Krahmer (2014) does not find differences in the perceptions of 188 news consumers depending on AI or human authorship attribution, but that differences appear in the perceptions of a group of 64 professional journalists.

The Product and the Process

The four P's of creativity, (Rhodes, 1961; Jordanous, 2014) distinguish between the 'Product' (the artefact produced by creativity), and the 'Process' (the set of actions taken leading to the production of an artefact). Although, we should keep in mind that different cultures and practices will focus more or less on the Product or the Process (Lubart, 2010), there is evidence that knowledge of, or even belief in, the production process of an artefact and the context in which it is produced is an important evaluative criterion for the resulting artefact (Chamberlain et al., 2018).

Compelling evidence of such a phenomenon is provided by Davies (2003), showing that one would not appreciate or value in the same way an original piece of art, truly novel for its time, reaching new frontiers of craftsmanship, and a newly made replica of it (of whatever fidelity), as has been shown by numerous cases of forgery (Bowden, 1999). Wolz and Carbon (2014) test this by showing participants artworks labelled as original or copies, showing that the alleged authenticity has a major impact on art appreciation. Another example is provided for music by Canonne (2018) who conducts a qualitative study where musicians listen to the same audio recording of a duet, but are either told they are listening to a composition or an improvisation. The interviews show that musicians' experience of the piece is very different in each condition, focusing on different aspects of the music, listening more to the acoustical features and overall structure when they believe they are listening to a composition, and listening more to the relational process and the interactions between instruments when they believe they are listening to an improvisation.

Related more directly to bias about the evaluation of Computational Creativity, research in neuropsychology (Steinbeis and Koelsch, 2009) shows that believing that an artefact is human-made (as opposed to AI-generated) activates areas of the cortex reported for mental state attribution, indicating that participants are engaging with the process and intentions of the alleged human artist. Moreover, as described in our survey, Chamberlain et al. (2018) discusses a study where participants observing the drawing process of a robot in a museum setting for as long as they want show a change in the parity of their assessments. This raises questions on the importance of audience engagement (Candy and Bilda, 2009) with the process and the product in this kind of study.

Safeguards and future directions

Considering the importance of socio-cultural contexts, we propose reframing the question of bias and Computational Creativity in order to better take context into account. Inspired by Lincoln and Guba (1985); Li (2004), we propose safeguards for future studies about bias in evaluating Computation Creativity:

- *Use thick descriptions:* In order to compare the research context of a study with those of other studies, we recommend using thick descriptive information regarding the methodology and the context. In particular, as many detailed information should be given regarding the stimuli used, the participants of the study, their relationship to the stimuli's style/topic, where and how the study was conducted, and other relevant contextual information (Ponterotto, 2006).
- *Refrain from generalizing:* As social and behavioural phenomena are bound by their specific contexts, we advise refraining from making generalizations about the results of a study about bias outside of the context studied even when results appear to be "statistically significant".
- *(Near-)Natural situation:* We strongly advise to closely align the research context with the artistically-relevant environment. A study should minimize external interference or changes that could be introduced as a result of the research. This applies to both the content of stimuli – which should be as 'natural' as possible regarding the style/topic – and the environment in which the stimuli are observed.
- *Triangulation:* We advise using triangulation as a mean of verifying both their data and interpretations. This involves using multiple sources of data, as well as different evaluation methods. Various data collection methods could be used such as surveys and interviews. One way of doing this is asking participants about their strategies for discriminating between human and AI authorship, as done by Déguernel, Sturm, and Maruri-Aguilar (2022); Chamberlain et al. (2018); Hitsuwari et al. (2023)

What does this entail for future research in Computational Creativity? First, regarding the evaluation of creative systems (Lamb, Brown, and Clarke, 2018), if biases need to be taken into account, then it means that a specific study regarding the specific context (or as close as possible) should be conducted, as studies conducted in different socio-cultural contexts may yield results that are irrelevant to the target domain. Therefore, Computational Creativity evaluation could actually become a great experimental ground, like in Norton, Heath, and Ventura (2015) or Hitsuwari et al. (2023), for better understanding the whys and wherefores of *algorithmic aversion* and where *algorithmic appreciation* arises in the scope of creativity, as they offer a large variety of applications with their respective socio-cultural contexts. Second, the results from all the studies presented in our survey show that questions around the existence of *algorithmic aversion* as a general direction of research in computational creativity, might not be the correct framing for future work. Instead, the field might focus more of the "what? where? when? who? and how?" of such biases, as this will lead to a better understanding of the impact of creative systems in society and help to lead more informed discussions in regard to AI ethics (Holzapfel, Jääskeläinen, and Kaila, 2022).

Acknowledgments

This work was supported by the grant ERC-2019-COG No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism.

References

Aljanaki, A. 2022. Attitude towards and evaluation of computer-generated music in music listeners and musicians. In *Proc. of the Conference on AI Music Creativity*.

Ariza, C. 2009. The interrogator as critic: The Turing test and the evaluation of generative music systems. *Computer Music Journal* 33(2):48–70.

Bowden, R. 1999. What is wrong with an art forgery?: An anthropological perspective. *The Journal of aesthetics and art criticism* 57(3):333–343.

Brieber, D.; Nadal, M.; Leder, H.; and Rosenberg, R. 2014. Art in time and space: Context modulates the relation between art experience and viewing time. *PloS one* 9(6).

Broussard, M.; Diakopoulos, N.; Guzman, A. L.; Abebe, R.; Dupagne, M.; and Chuan, C.-H. 2019. Artificial intelligence and journalism. *Journalism & Mass Communication Quarterly* 96(3):673–695.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; et al. 2020. Language models are few-shot learners. *CoRR* abs/2005.14165.

Candy, L., and Bilda, Z. 2009. Understanding and evaluating creativity. In *Proc. of the 7th ACM Conference on Creativity and Cognition*, 497–498.

Canonne, C. 2018. Listening to improvisation. *Empirical Musicology Review* 13(1–2).

Chamberlain, R.; Mullin, C.; Scheerlinck, B.; and Wagemans, J. 2018. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity and the Arts* 12(2):177–192.

Clerwall, C. 2014. Enter the robot journalist: users' perceptions of automated content. *Journalism Practice* 8(5).

Dahlig, E., and Schaffrath, H. 1993. Komputerowa symulacja melodii ludowych. *Eksperyment Muzyka* 38(3–4).

Dahlig, E., and Schaffrath, H. "1998". Judgments of human and machine authorship in real and artificial folksongs. *Computing in Musicology*.

Darda, K. M., and Cross, E. S. 2022. The role of expertise and culture in visual art appreciation. *Scientific Reports* 12(1):10666.

Davies, S. 2003. Ontology of art. In Levinson, J., ed., *The Oxford Handbook of Aesthetics*. Oxford University Press.

Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. of Exp. Psychology: General* 144(1).

Dubnov, S.; Burns, K.; and Kiyoki, Y. 2016. Cross-cultural aesthetics: analyses and experiments in verbal and visual arts. In *Cross-Cultural Multimedia Computing: Semantic and Aesthetic Modeling*. Springer. 21–41.

Déguernel, K.; Sturm, B. L. T.; and Maruri-Aguilar, H. 2022. Investigating the relationship between liking and belief in AI authorship in the context of Irish traditional music. In *Proc. of CREAI: Workshop on Artificial Intelligence and Creativity*.

Eigenfeldt, A.; Burnett, A.; and Pasquier, P. 2012. Evaluating musical metacreation in a live performance context. In *Proc. of ICCC*.

Fischinger, T.; Kaufmann, M.; and Scholtz, W. 2018. If it's mozart, it must be good? the influence of textual information and age on musical appreciation. *Psychology of Music* 48(4):579–597.

Flick, C., and Worrall, K. 2022. The ethics of creative AI. In Vear, C., and Poltronieri, F., eds., *The Language of Creative AI: Practices, Aesthetics and Structures*. Springer International Publishing. 73–91.

Flôres, R. G., and Ginsburgh, V. A. 1996. The Queen Elisabeth musical competition: How fair is the final ranking? *The Statistician* 45(1):97–104.

Friedman, R. S., and Taylor, C. L. 2014. Exploring emotional responses to computationally-created music. *Psychology of Aesthetics, Creativity and the Arts* 8(1):87–95.

Gangadharbatla, H. 2021. The role of AI attribution knowledge in the evaluation of artwork. *Empirical Studies of the Arts* 40(2).

Graefe, A.; Haim, M.; Haarmann, B.; and Brosius, H.-B. 2018. Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism* 19(5):595–610.

Greasley, A., and Lamont, A. 2016. Musical preferences. In Hallam, S.; Cross, I.; and Thaut, M., eds., *The Oxford Handbook of Music Psychology*. Oxford University Press.

Hagtvedt, H.; Patrick, V. M.; and Hagtvedt, R. 2008. The perception and evaluation of visual art. *Empirical studies of the arts* 26(2):197–218.

Herremans, D.; Chuan, C.-H.; and Chew, E. 2018. A functional taxonomy of music generation systems. *ACM Computing Surveys* 50(5).

Hitsuwari, J., and Nomura, M. 2022. How individual states and traits predict aesthetic appreciation of haiku poetry. *Empirical Studies of the Arts* 40(1):81–99.

Hitsuwari, J.; Ueda, Y.; Yun, W.; and Nomura, M. 2023. Does human–AI collaboration lead to more creative art? aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* 139(107502).

Holzapfel, A.; Jääskeläinen, P.; and Kaila, A.-K. 2022. Environmental and social sustainability of creative AI. In *Generative AI and CHI Workshop*.

Hong, J.-W., and Curran, N. M. 2019. Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15(2).

Hong, J.; Fischer, K.; Ha, Y.; and Zeng, Y. 2022. Human, I wrote a song for you: An experiment testing the influence of machines' attributes on the AI-composed music evaluation. *Computers in Human Behavior* 131(107239).

Hong, J.-W.; Peng, Q.; and Willians, D. 2021. Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society* 23(7):1920–1935.

Hong, J.-W. 2018. Bias in perception of art produced by artificial intelligence. In *Human-Computer Interaction. Interaction in Context*, 290–303.

Huang, R.; Sturm, B. L. T.; and Holzapfel, A. 2021. De-centering the West: East Asian philosophies and the ethics of applying artificial intelligence to music. In *Proc. of ISMIR*.

Hämäläinen, M., and Alnajjar, K. 2019. Let's FACE it. Finnish poetry generation with aesthetics and framing. In *Proc. of the 12th International Conference on Natural Language Generation*, 290–300.

Jago, A. S. 2019. Algorithms and authenticity. *Academy of Management Discoveries* 5(1).

Jordanous, A. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proc. of ICCC*.

Kirk, U.; Skov, M.; Hulme, O.; Christensen, M. S.; and Zeki, S. 2009. Modulation of aesthetic value by semantic context: An fMRI study. *NeuroImage* 44(3):1125–1132.

Koelsch, S.; Vuust, P.; and Friston, K. 2019. Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences* 23(1):63–77.

Kroger, C., and Margulis, E. H. 2016. "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music* 45(1):49–64.

Lahdelma, I., and Eerola, T. 2020. Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension. *Scientific Reports* 10:8693.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys* 51(2).

Lauring, J. O.; Pelowski, M.; Forster, M.; Gondan, M.; Ptito, M.; and Kupers, R. 2016. Well, if they like it... effects of social groups' ratings and price information on the appreciation of art. *Psychology of Aesthetics, Creativity, and the Arts* 10(3).

Leder, H., and Nadal, M. 2014. Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–developments and challenges in empirical aesthetics. *British journal of psychology* 105(4):443–464.

Lermann Henestrosa, A.; Greving, H.; and Kimmerle, J. 2023. Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Computers in Human Behavior* 138(107445).

Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; et al. 2022. Competition-level code generation with alphacode. *Science* 378(6624):1092–1097.

Li, D. 2004. Trustworthiness of think-aloud protocols in the study of translation processes. *International Journal of Applied Linguistics* 14(3):301–313.

Lincoln, Y., and Guba, E. G. 1985. *Naturalistic inquiry*. Sage.

Liu, B., and Wei, L. 2019. Machine authorship in situ: effect of news organization and news genre on news credibility. *Digital Journalism* 7(5):635–657.

Logg, L. M.; Minson, J. A.; and Moore, D. A. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Longoni, C.; Fradkin, A.; Cian, L.; and Pennycook, G. 2022. News from generative artificial intelligence is believed less. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106.

Lubart, T. I. 2010. Cross-cultural perspectives on creativity. In *The Cambridge Handbook of Creativity*. Cambridge University Press. 265–278.

Maher, T. F. 1976. "need for resolution" ratings for harmonic musical intervals: A comparison between indians and canadians. *J. of Cross-Cultural Psychology* 7(3).

Mahmud, H.; Islam, A. N.; Ahmed, S. I.; and Smolander, K. 2022. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change* 175(121390).

McDermott, J. H.; Schultz, A. F.; Undurraga, E. A.; and Godoy, R. A. 2016. Indifference to dissonance in native amazonians reveals cultural variation in music perception. *Nature* 535(7613):547–550.

Moffat, D. C., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. *Assessment* 13(11).

Moura, F. T., and Maw, C. 2021. Artificial intelligence became Beethoven: how do listeners and music professionals perceive artificially composed music. *Journal of Consumer Marketing* 38(2):137–146.

North, A.; Hargreaves, D. J.; and Hargreaves, J. J. 2004. Uses of music in everyday life. *Music perception* 22(1).

Norton, D.; Heath, D.; and Ventura, D. 2015. Accounting for bias in the evaluation of creative computational systems: an assessment of DARCI. In *Proc. of the 6th International Conference on Computational Creativity*, 31–38.

Obermeier, C.; Menninghaus, W.; Von Koppenfels, M.; Raettig, T.; Schmidt-Kassow, M.; Otterbein, S.; and Kotz, S. A. 2013. Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in psychology* 4.

Orr, M. G., and Ohlsson, S. 2005. Relationship between complexity and liking as a function of expertise. *Music perception* 22(4):583–611.

Pasquier, P.; Burnett, A.; Gonzalez Thomas, N.; Maxwell, J. B.; Eigenfeldt, A.; and Loughin, T. 2016. Investigating listener bias against musical creativity. In *Proc. of the 7th International Conference on Computational Creativity*.

Pearce, M. T. 2015. Effects of expertise on the cognitive and neural processes involved in musical appreciation. In Huston, J. P.; Nadal, M.; Mora, F.; Agnati, L. F.; and Cela-Conde, C. J., eds., *Art, aesthetics and the brain*. Oxford University Press. 319–338.

Pease, A., and Colton, S. 2011. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *Proc. of the AISB Symposium on AI and Philosophy.*, 15–22.

Ponterotto, J. G. 2006. Brief note on the origins, evolution, and meaning of the qualitative research concept 'thick description'. *The Qualitative Report* 11(3):538–549.

Ragot, M.; Martin, N.; and Cojean, S. 2020. AI-generated vs. human artworks. a perception bias towards artificial intelligence? In *CHI Conference on Human Factors in Computing Systems*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *Proc. of the 37th International Conference on Machine Learning*, 8821–8831.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models. *CoRR* abs/2112.10752.

Shank, D. B.; Stefanik, C.; Stuhlsatz, C.; Kacirek, K.; and Belfi, A. M. 2022. AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied* 28.

Steinbeis, N., and Koelsch, S. 2009. Understanding the intentions behind man-made products elicits neural activity in areas dedicated to mental state attribution. *Cerebral Cortex* 19(3):619–623.

Strapparava, C., and Stock, O. 2011. Computational humour. In Cowie, R.; Pelachaud, C.; and Petta, P., eds., *Emotion-Oriented Systems*. Springer. 609–634.

van der Kaa, H. A. J., and Krahmer, E. J. 2014. Journalist versus news consumer: The perceived credibility of machine written news. In *Proc. of the Computation+Journalism conference*.

Waddell, T. 2018. A robot wrote this?: how perceived machine authorship affects news credibility. *Digital Journalism* 6(2):236–255.

Winston, A. S., and Cupchik, G. C. 1992. The evaluation of high art and popular art by naive and experienced viewers. *Visual Arts Research* 18(1).

Wolz, S. H., and Carbon, C. C. 2014. What's wrong with an art fake? cognitive and emotional variables influenced by authenticity status of artworks. *Leonardo* 47(5):467–473.

Wu, Y.; Mou, Y.; Li, Z.; and Xu, K. 2020. Investigating american and chinese subjects' explicit and implicit perceptions of AI-generated artistic work. *Computers in Human Behavior* 104(106186).

4. Image Generation and Processing, Sound and Music

Emojinator: An Emoji Generator to Represent Emotions

Marzia Bilwani and Anna Jordanous
School of Computing
University of Kent
Canterbury, UK
marziabilwani@gmail.com, a.k.jordanous@kent.ac.uk

Abstract

In recent years, emojis have become a key part of computer-mediated communication (CMC). This stems from the fact that they act as nonverbal cues which are difficult to convey when communicating using simple text. Just as a picture is worth a thousand words, the same can be said for an emoji. In this paper, we present a creative system, *Emojinator*, that generates emojis using visual blending to represent a diverse range of emotions. Unlike previous emoji generation work, fuzzy logic was incorporated to enable Emojinator to make decisions. A user study was conducted to evaluate the output, along with the creative tripod method to assess Emojinator's creativity. The results from the survey show that for more than half the emojis, at least 50 percent of participants agreed or strongly agreed that the emoji represented the stated emotion. Evaluation through the creative tripod method showed that the system is skillful and imaginative but could be more appreciative. Therefore, further refinement may be needed to make the system more creative. However, the success of this novel approach to emoji creation opens up new directions for future work.

Introduction

In recent years, the use of emojis has increased rapidly. Literally meaning "picture-word" in Japanese, their popularity in written language can be seen in the rise in the number of emoji-related tools such as search-by-emoji and emoji replacement or prediction features. The emoji language has also been proposed as the fastest-growing language in the UK (Doble 2015), referencing how a large proportion of 18 to 25-year-olds find it easier to express their feelings through emojis instead of through text. Emojis gained more popularity after the Unicode standard incorporated them, and after Apple added the emoji keyboard to iOS in 2011, with new emojis released every year since (Dimson 2015). Since then, their number has increased continuously with the addition of new characters in Unicode, comprising not just faces but also pictographs depicting vehicles, buildings, food, drinks, activities like dancing and running, and animals and plants (Pavalanathan and Eisenstein 2015).

The increasing number of emojis does not indicate a corresponding increase in emojis with visual representations of emotion. Based on recent numbers, there are a total of 3633 emojis in the Unicode Standard; out of these 157 are single-character smileys representing emotions (EmojiList 2022). Consequently, developing new emojis to represent a greater number of emotions seems to be a laborious task. With the growth of computational creativity, it therefore makes sense to delegate this task to a creative system.

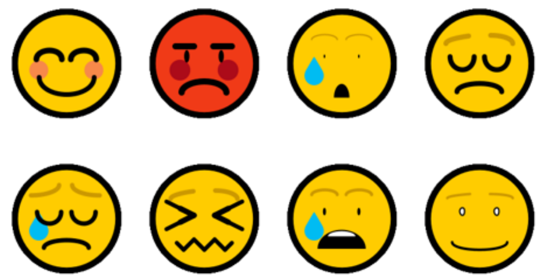


Figure 1: Examples of emojis generated by Emojinator. From left to right and top to bottom: *Joyous, Bitter, Anxious, Disappointed, Dissatisfied, Miserable, Worried, Satisfied*.

In this paper, we develop a system that uses *visual blending* to generate new emojis. Visual blending, based on the idea of the Conceptual Blending (CB) theory (Fauconnier and Turner 2002) is the creation of new visuals, such as images, by combining at least two current ones (Cunha, Martins, and Machado 2018b). The *Emojinator* system creates new emojis by blending features from an emoji (Figure 1). The system uses fuzzy logic to make decisions on what kind of features an emoji should have to depict a particular emotion. To evaluate the results and understand how the system can be improved, a user study was conducted. We also evaluate the creativity of the system using the creative tripod (Colton 2008). Before discussing *Emojinator* further, it is important to explore the significance of emojis to understand why this area is important.

Importance of Emojis

It is a well-known fact that emojis have become an important part of digital communication. Major technological companies have realised their significance as well and have taken several steps to incorporate emojis in their systems. Alongside the business importance of emojis, there are psychological, sociological and linguistics-related aspects to emojis.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Psychological Aspects Before the introduction of emojis, emoticons were used to display emotions in communication through texting, email, and other forms of computer-mediated communication. Emoticons, unlike emojis, are letters, punctuation marks or numbers that usually represent an emotion, for instance, a smiling face would be ':)'. As computer-mediated communication is devoid of nonverbal cues, the primary objective of emoticons was to translate emotions to convey facial expressions (Walther and D'Addario 2001). It has been found that similar parts of the brain are activated when a person sees a smiling emoticon or emoji, as when they see someone smiling in real life (Churches et al. 2014).

This function has developed with time with the growth of online systems and emojis. In a study on how emojis impact emotional communication and information processing, it was found that understanding of verbal messages and processing speed were found to have been improved by adding emojis (Boutet et al. 2021). The results of this study thus supported use of emojis, especially positive ones, to enhance communication.

Sociological Aspects In face-to-face settings, nonverbal cues help in communicating information which impacts our perception of other people and our behaviours towards them as well (Stewart et al. 2012). For instance, individuals who smile more frequently may be considered 'warmer' (Wang et al. 2017). A similar effect is seen when emoticons and emojis are used.

Linguistics-related Aspects In recent years, research has extensively examined the role of emoticons in communication. These symbols are believed to convey emotions and thoughts by mimicking nonverbal cues (Crystal 2006). Nonverbal cues are usually the main piece of information processed by the brain and when an emoticon or emoji is seen, it is identified as an emotional interaction (Yuasa, Saito, and Mukawa 2011). Emojis are perceived as not words but emotional information as they help to articulate the tone of voice and gestures which are usually only possible when people are communicating vocally.

Consequently, emojis play a significant part in helping people express their emotions and helping others in understanding them. It then becomes crucial to understand emotions and how they can be modelled to represent emotions.

Computational Modelling of Emotions

Emotions are one of the key significant unconscious mechanisms that affect human behaviours, decision making and attention (Phelps 2006). As there are several elements and facets which underly emotions, they can be approached from various perspectives. The multi-faceted nature of emotions has resulted in them being the focus of study in various disciplines such as neuroscience, cognitive informatics, psychology, philosophy and computer science (Wang 2007b). This multidisciplinary study has led to the development of several computational, cognitive and theoretical models.

While there are several theories of emotions, Ekman's model, Wang's Hierarchical model of emotions and Russell's circumplex model will be discussed as these are some of the main theories in this field.

Ekman's Model

One of the most well-known theories of emotions is Ekman's model of six basic emotions comprising sadness, surprise, fear, happiness, disgust and anger (Ekman 1999), based on different facial expressions. However, these are usually not used in the development of cognitive computational models of emotions (Rodríguez, Ramos, and Wang 2012). According to Cohen (2005), the model of basic emotions does not have the conceptual room to consider emotional experiences and therefore, is not an adequate theory of emotion. The fact that it depicts only six emotions also limits its usability for this system since these emotions are already depicted in emojis currently.

Wang's Hierarchical Model of Emotions

A hierarchical model of emotions was developed by Wang (2007a). In this model, human emotions were classified into two categories: unpleasant and pleasant. Emotions in the two categories can be further classified into five levels based on the intensity of subjective feelings where every level consists of a pair of pleasant and unpleasant emotions. While the hierarchical model of emotions is wider in scope in comparison to Ekman's model, its focus on the link between emotions, attitudes and motivations makes it difficult to apply this model to this study; the primary objective here is to use emojis to depict emotions, not to study the underlying motivations and attitudes behind emotions.

Russell's circumplex model

In Russell's circumplex model of affect, emotions are modelled spatially in which eight variables are plotted on a two-dimensional graph (Russell 1980). The dimensions used in this graph are:

- **Valence:** the extent to which an emotion is positive or negative. E.g. delighted is a positive valence emotion in Russell's model, while sad is a negative valence emotion.
- **Arousal** the intensity of emotion. It ranges from calm (low) to excited (high).

Scherer's update to the Russell model A problem which arises in Russell's model is how certain emotional states may fall under a similar area of the two-dimensional space – for instance, both angry and tense would have negative valence and high arousal. In such a situation, verbal labels help to identify key components of the stimulating event and the integrated interpretation of reaction patterns (Scherer 2005). Using emotional labels, and incorporating goal conduciveness, coping potential, and appraisal dimensions with the strongest impact on emotions, Scherer (2005) superimposed a two-dimensional structure on Russell's model with various emotion terms (indicated with a +, lower-case words). This addition by Scherer led to a wider variety of emotions being represented in this model.

Paltoglou and Thelwall (2013) used Scherer's model for measuring the emotional content of blog posts. During this

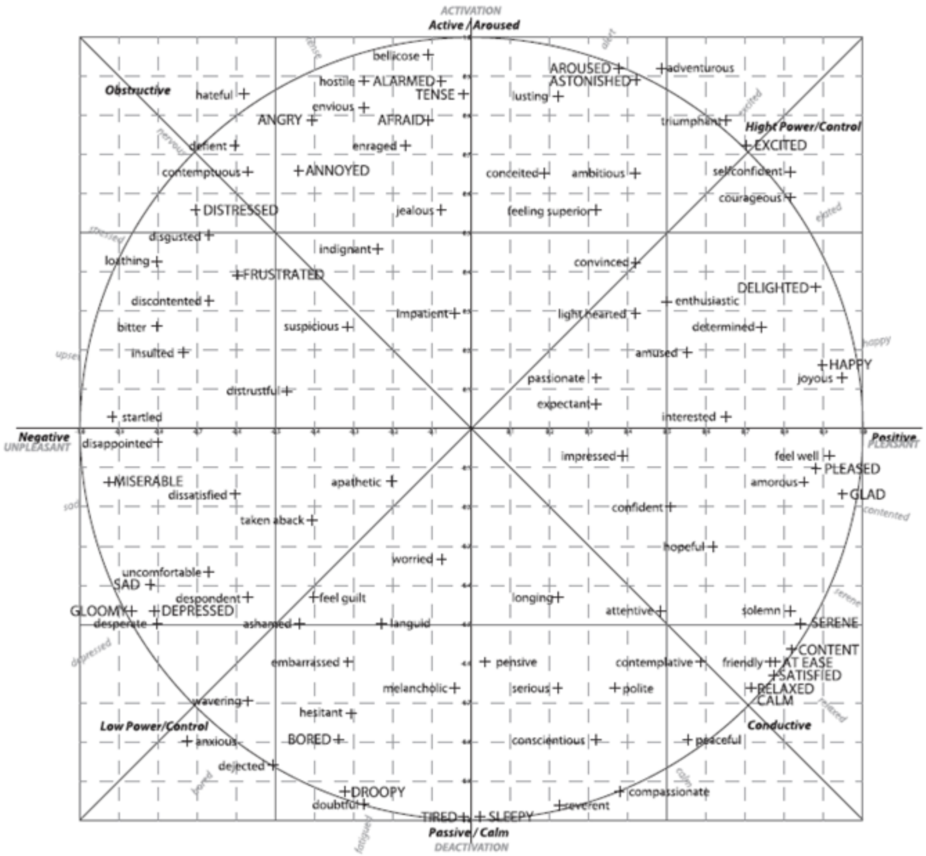


Figure 2: Scherer's updates to Russell's model converted into quantitative data. Source: (Paltoglou and Thelwall 2013). The diagram has been used with permission from the authors. Upper-case notation represents the terms that were used by Russell (1980).

application, they converted the graphical data into quantitative data to use in their study. This conversion makes the model quite useful for implementation in this project. The model contains a wide and diverse range of emotions across the full range of dimensions of valence and arousal – a total of 97 emotions. Therefore, the Russell model, as updated by Scherer (2005) and Paltoglou and Thelwall (2013) was used in the development of this system (Figure 2).

The Approach

Computational creativity work on emoji generation is mostly focused on two approaches: Generative Adversarial Networks (GANs) which have been used to replicate existing emojis (Radpour and Bheda 2017; Puyat 2017) (Figure 3) and visual blending (Cunha, Martins, and Machado 2018a; 2018b; Cunha et al. 2019; 2020) (Figure 4). As GANs have not been able to achieve the same level of sophistication, visual blending has been used here. In this section, the process of visual blending is discussed, followed by an explanation of the key components of this system.

Visual Blending

Visual blending, based on the idea of the Conceptual Blending (CB) theory (Fauconnier and Turner 2002) is the creation of new visuals, such as images, by combining at least two current ones (Cunha, Martins, and Machado 2018b). There are several examples of visual blending. The two relevant ones are: character blending for Pokémon (name and image) in which mappings exist between attributes, such as colour and shape and type, resulting in a new type of Pokémon (Liapis 2018). In addition, the X-Faces system generates new faces by merging different face parts to enhance data augmentation in face detection (Joao Correia and Machado 2016). Similar work has been done with emojis as well.

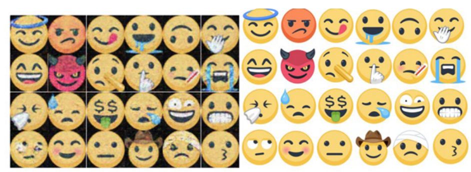


Figure 3: Emojis generated by conditioning the network (left) with actual emojis (right). Source: (Radpour and Bheda 2017). The image has been used here with permission from the authors.

Emoji Generation using Visual Blending Before emojis, emoticons were used and the ease with which individual parts of an emoticon could be changed, for instance, chang-

Figure 4: Blends for *peace accord, car factory, security, house, market depression, health risk* and *airline bureaucracy*. Source: (Cunha et al. 2020). The image has been used here with permission from the authors.

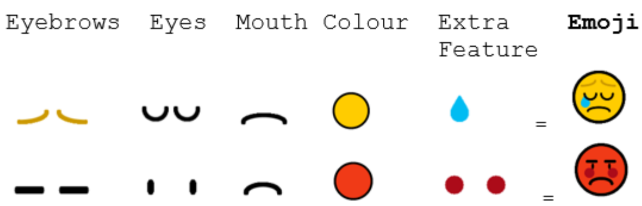


Figure 5: Examples of two emojis created by blending features

ing a bracket from ")" to "(" to make an emoticon ":(" could mean a sad face, has led to the development of different blending approaches to generate emojis.

In 2016, the Unicode Consortium decided to introduce the ZWJ (Zero-Width-Joiner) method which mainly consisted of an invisible character to describe the combination between two characters (Abbing, Pierrot, and Snelting 2017). A key example of emoji generation using visual blending is the *Emojinating* system that blends existing emojis to generate new ones to enhance creativity and assist in the idea generation process (Cunha, Martins, and Machado 2018a). This system has a wide range of applications such as helping in idea generation and designing icons (Cunha, Martins, and Machado 2018a). Another emoji generator, *Emojimoji*, generates new emojis by randomly combining two existing ones (EmojiMoji 2022).

Cunha et al. (2020) further worked in this area by assessing the Emojinating system to gauge the suitability of this approach for the visual depiction of concepts. However, the focus of Emojinating is quite different from this system in the sense that it does not specifically focus on emotions and mainly used emojis as a case study to gauge the effectiveness of visual blending for concept representation. By focusing on emotions, our system addresses a gap in existing research. The visual representation of emotions through emojis interlinks with the psychological aspect of emojis helping in communication and enabling people to express emotions more easily.

Visual blending was also used in this project because of similarities between existing emojis. For example, most emojis had common features such as eyes, face, mouth, and eyebrows and were yellow in colour. There were different types of each feature. For instance, if we focused on eyes,there are oval, smiling and x-shaped eyes. This similarity meant that once common features were identified they could be used in different emojis based on the emotion.

However, the question then arose of how exactly should visual blending be done? What can be blended to create a new emoji and what tools should be used?

Features of an Emoji To identify the similarities between emojis, we mapped the existing emojis to the circumplex model using *Emojipedia* (Emojipedia 2022) as a guide (see Figure 6). This helped to make links in features between different emojis. Blending these features, would therefore result in the generation of new emojis (see Figure 5).

The artwork selected for emojis was the one that is visible in a browser. This was to ensure ease and standardization as the artwork varies according to each platform. For face colour, yellow and red were selected. Yellow is used in most emojis and red was chosen as other emotions in the second quadrant, such as *jealous* and *indignant* had a similar meaning associated with them as *enraged*. While the *disgusted* emoji has a green colour face, this was not incorporated as disgust has a distinct relationship with the colour green and no other emotion in the model has a similar meaning.

Fuzzy logic was then used to represent the overlap across various features to represent different emotions.

Computational Tools and Fuzzy Logic

Py5 (py5coding 2022), a new version of *Processing* for Python, was used. Py5 is a widely used software sketchbook that is used to create images with code. To ensure that the system is autonomous and is not completely following the rules set by a human being, fuzzy logic was incorporated. Fuzzy logic can be used in situations where there is a possibility for imprecision (Zadeh 1996). This was needed in this system since it is difficult to classify emotions based on just crisp logic. The below examples demonstrate the difference between the two logical processes using emotions:

- **Crisp logic:** If Sarah passes her dissertation (gets a mark above 50), she will be happy, otherwise she will be sad.
- **Fuzzy logic:** The degree to which Sarah is sad or happy will depend on her overall mark instead of being binary. If she scores above 70, she will probably be ecstatic but there is a small possibility she might be sad as she wanted a 90. Otherwise, if she scores a 50, she will be sad, but also relieved about passing her final module.

Fuzzy logic provided a way to convert the two-dimensional data on emotions (valence and arousal) to a one-dimensional space, while also making the system *artificially intelligent*. An agile software development process was also followed to allow more room for flexibility. The objective of this was to give the system *creative* freedom.

Fuzzy Logic Architecture The first step in fuzzy logic architecture is *fuzzification* in which a crisp input value is used to determine the extent to which the input belongs to a fuzzy set (Guo and Wong 2013). To depict fuzzy sets graphically, trapezoidal membership functions (MF) were used. Two linguistic variables were defined based on the circumplex model – valence and arousal, which would determine the output, that is, emotional state. A trapezoidal membership function was used, since it covered a larger area.

Five membership functions were defined for valence and arousal: very low, low, medium, high and very high (Figure 6). This was predominantly to capture a diverse range

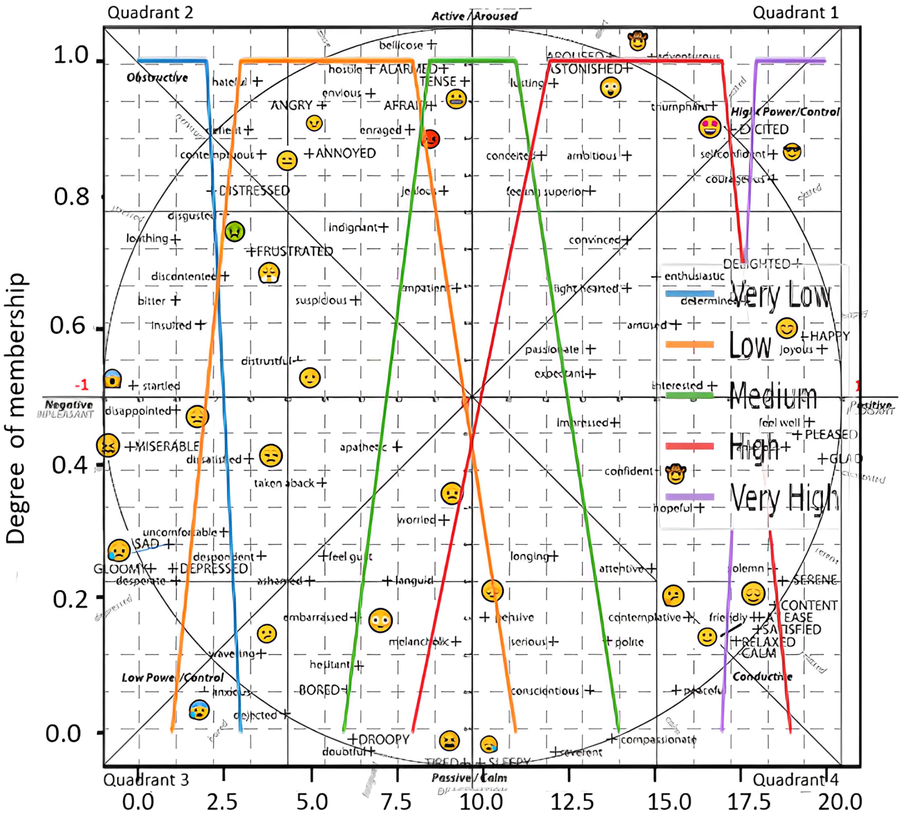


Figure 6: Fuzzy logic membership functions for Valence and existing emojis mapped on to the circumplex model

of emotions from the circumplex model more effectively. A neutral MF was incorporated to effectively represent emotions such as tired and sleepy which have a very low valence on the circumplex model but lie on the negative and positive sides respectively of the x-axis due to the slightly different connotations of the two words. A similar process was followed for output fuzzy sets for emotional state: very unpleasant, unpleasant, neutral, pleasant and very pleasant.

After fuzzification, fuzzy rules had to be defined. For simplicity, these were split into various combinations of valence and arousal and what emotional state they would lead to, using the fuzzy operators AND, NOT and OR which were available in the Simpful library (Spolaor et al. 2020). The Mamdani Inference system was used as it is better suited to human inputs, with a more interpretable rule base, making it more appropriate for this project. On the other hand, the Sugeno inference system is better suited to mathematical analysis and makes use of a singleton output MF that is a linear function (mathworks.com 2022).

Once an emotional state value was determined, it was assigned a positive or negative sign based on the quadrant it fell under. For example, in quadrant 2, the overall output value would be negative since valence is negative, while arousal is positive, and in quadrant 1, the overall value would be positive as both valence and arousal values are positive.

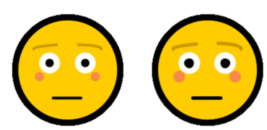


Figure 7: Emojis for *embarrassed* (left) and *felt guilt* (right)

Rules and Uniqueness

Once an emotional state value is calculated, if-then rules use these values to decide what feature every emoji should have.

To ensure that every emoji looked unique, a weight parameter, determined by the arousal value was added to the features (eyes, mouth etc). This would impact stroke thickness and the qualities of features depending on their attributes. For example, in Figure 7, stroke thickness for mouth and eyebrows, and diameter for eyes and flushed cheeks are slightly different for the two emojis. We recognize that some of the emojis look very similar and this is a limitation of this work.

How it Works A user has two options to create an emoji:

- Entering valence and arousal values
- Selecting an emotion from the dropdown box (a list of all the emotions from the circumplex model).

Figure 8: Emojis used in Section 1 of the survey. From left to right and top to bottom: *Light Hearted, Lusting, Passionate, Frustrated, Tense, Insulted, Taken Aback, Droopy, Embarassed, Contemplative, Confident* and *Serious*

Based on the values entered between the ranges of -1 to 1 or selection of the emotions, an emoji is generated.

Results and Discussion

In this section, we present and discuss the results generated. Following this, the system is evaluated using the *creative tripod* approach.

User Testing

Emojis are mostly used in computer-mediated communication to help individuals communicate more easily. Thus, it becomes essential to acquire responses from human participants to understand how they would be interpreting an emoji. A survey was created to get feedback from participants about the emojis generated by the *Emojinator* system. The survey was split into three sections.

In the first section, participants had to evaluate emojis based on how accurately they represented the stated emotion. A five-point Likert scale was used in which participants had to select their level of agreement: (1) Strongly agree (2) Agree (3) Neutral (4) Disagree and (5) Strongly disagree. A Likert scale was used, as ratings are ordinal values (Yannakakis and Martinez 2015); therefore, assigning relative values to emotions is a better approach than using absolute values due to the ordinal nature of emotions (Yannakakis, Cowie, and Busso 2018). At the end of the section, participants were asked to share any comments they had about the emojis to acquire qualitative feedback as well. A total of 12 emojis were used in this section with three from every quadrant of the circumplex model. They were selected randomly using a random word selector website (textfixer.com 2022) to avoid bias on our part in selecting emojis (Figure 8).

The second section involved a comparison of system-generated and existing emojis with the objective of understanding which emoji better represented the specific emotion. Here we compare to existing emojis, as we are treating existing emojis as the benchmark for this work. Participants were not told which ones were system-generated and which ones are the existing ones.

Finally, in the third section, participants were told which emojis were system-generated and which are the existing ones. They were again asked which emoji better represented the emotion mentioned and to provide a reason for their choice. This was to gauge if opinions had changed. The objective of adding a comment box was to get both quantitative and qualitative results to analyze the output of Emojinator. The same rendering of emojis was used for comparison with existing emojis to keep the results consistent.

Analysis of Responses A total of 45 responses were received in the survey.

Looking at Figure 9 shows that for 7 out of 12 emojis at least fifty percent of participants agreed or strongly agreed about how effectively an emoji represented a given emotion. The top-rated emojis were *embarrassed*, followed by *frustrated*, *light-hearted* and *insulted* emojis. In terms of the lowest ranked emoji, 69 percent of participants disagreed or strongly disagreed about how effectively the emoji for *passionate* depicted the emotion. This was followed by the *lusting*, *serious* and *droopy* emojis.

Out of the 12 emotions listed in this section, *frustrated, tense, embarrassed* and *contemplative* have existing emojis to represent them. Out of these, *embarrassed* closely resembles its existing counterpart. This could be a possible explanation for why the system-generated emoji for embarrassed ranked the highest out of all these emojis suggesting that participants were more used to existing emojis. There was also an interesting comment about how the frustrated emoji should have clenched teeth. This again proves to some extent that individuals are used to seeing existing emojis and this could potentially impact their opinions. However, participants did score *frustrated* (73 percent of participants agreeing or strongly agreeing) and *tense* (58 percent of participants agreeing or strongly agreeing) highly in terms of their effectiveness in representing their respective emotions despite them looking quite different from their existing counterparts. This suggested that *Emojinator* was producing emojis that represented emotions well.

The comment box at the end of this section also had some insightful feedback with comments about specific emojis such as *insulted* looking angry and *tense* looking sad. Another respondent mentioned how neutral was selected as a response as the emoji could represent a particular emotion but was more representative of another emotion. This showed there was some ambiguity regarding interpretation.

Comparison with Existing Emojis Four emotions were selected in this section, one from each quadrant – *happy*, *worried*, *enraged* and *pensive* using a random word selector (textfixer.com 2022). Users were not told which emojis were system-generated and which were the existing ones and had to decide which emoji better represented an emotion. Figure 10 shows the results from this section.

The results show overwhelming support for existing emojis with the only exception being *pensive*. In the next section, participants were told which emojis were system-generated and which were the existing ones and then asked to answer which depicted the emotion more effectively. There were slight changes in the results with more responses for Emojinator, suggesting possible bias from respondents to indicate their preference for system-generated emojis (Figure 11).

However, the overall trend remained the same. Since participants could give feedback in this section, several interesting insights emerged. The impact of eyebrows on how well an emoji depicted an emotion could be seen in the happy

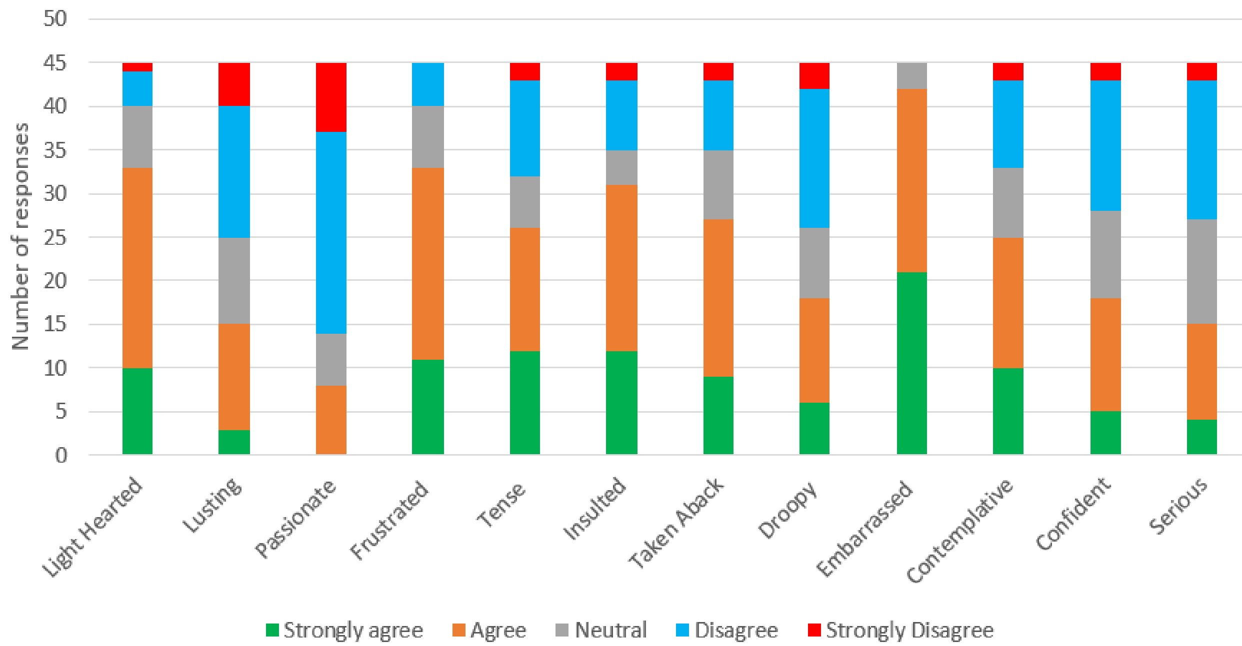


Figure 9: Section 1 Responses. The emoji for *embarrassed* was the best representation, while *passionate* had the worst

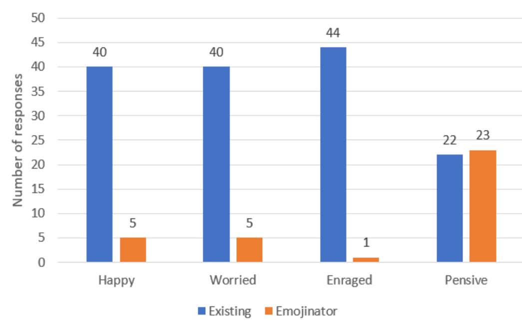


Figure 10: Section 2 Responses. *Pensive* was the only emoji with higher ratings for Emojinator output

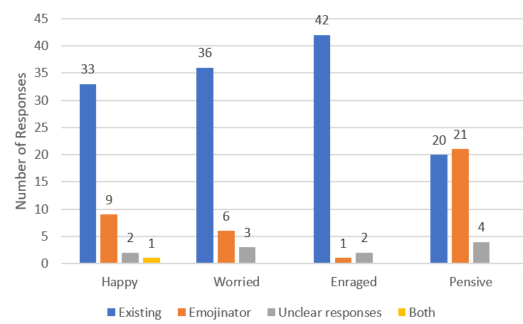


Figure 11: Section 3 Responses

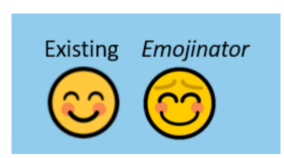


Figure 12: The *Happy* emoji

and enraged emojis. For the former, the addition of eyebrows resulted in comments such as how the smile seems 'strained' and how the existing emoji seems simpler in comparison since it does not have eyebrows (Figure 12). Similarly, the shape of the eyebrows in the enraged emoji led to comments such as how it looks sad because of the orientation of the eyebrows (Figure 13).

Results for the *pensive* emoji were better than the existing emoji since the existing one looks sadder according to the overall feedback. The Emojinator generated *pensive* emoji therefore also performed best in the final section.

Summary of Results To summarize the findings, more than half of the participants agreed or strongly agreed that 7 out of 12 emojis effectively depicted the stated emotion. When comparing system-generated to existing emojis, they rated the existing emojis higher for 3 out of 4 emojis. While these results offer insights into how well participants interpreted the system-generated emojis, interpretation may always be subjective. For instance, participants said they liked the Emojinator-generated emojis more once they found out which ones are system-generated. Besides this, participants could have been more used to emoji renderings on different operating systems, such as Apple and Google smartphones, which have been found to affect how people interpret the same emoji (Miller et al. 2017). A person's interpretation of emojis is also impacted by age (Koch, Romero, and Stachl 2022; Jaeger et al. 2018). Individuals above the age of 30

Figure 13: The *Enraged* emoji

tend to interpret emojis more literally, compared to younger people who interpret them more customarily (Herring and Dainas 2020). Therefore, evaluating creativity, in this case, emoji interpretation is a subjective area. However, trends in the above data do indicate that some form of consensus does exist among participants which will be useful in future work.

Evaluation

Evaluating creativity of Emojinator Evaluating creativity of a creative system can be complex due to the varying definitions of creativity. We used the creative tripod approach (Colton 2008), evaluating systems not just on output but also how they produce artefacts, using three principles:

- **Skillful:** The system is skillful since it can generate emojis. However, defining the level of skillfulness is difficult. Based on the way emojis are generated, the system is currently limited to just features such as eyes, eyebrows etc that have been written in code. Skillfulness can perhaps be enhanced by automating this process.
- **Appreciative:** At this stage, Emojinator has to be told what emotion to represent in an emoji. This can be achieved by a user entering a specific word or inputting valence and arousal values. Therefore, the software knows it has to generate an emoji based on the emotion or valence and arousal values it receives. However, the system does not fully know the value of its artwork. This is something which can be improved upon.
- **Imaginative:** The system is also imaginative in the sense that it makes every emoji unique by altering its features based on the arousal value which in turn determines the weight value. By doing this, it is generating unique and novel emojis. A limitation is that some of the emojis, however, look very similar (see Figure 7). By using fuzzy logic, the system has some autonomy in its decision-making. However, the software is still generating emojis based on rules that use the emotional state value – it is essentially *taught* how to be imaginative. Moreover, it could be more imaginative by adding a machine vision component, that could perhaps detect emotions that a person expresses and generate them in the form of an emoji.

User study and Results A limitation of the user study is that only four emojis were chosen to compare with existing emojis in sections 2 and 3. To better represent the diverse emotions in the circumplex model, more emojis could have been added. However, as it was an online survey, adding more emojis would have made the survey longer. This could have led to a decrease in the respondents' attention and response quality, as online survey respondents generally have shorter attention spans (Fricker and Schonlau 2002).

While 50 percent of participants agreed about representation for 7 emojis, the remainder did not fully associate the emojis with the stated emotions. A ranking approach that would have helped the user identify the second or third best emotion for a particular emoji would have been better. The distance between two emotions on the circumplex model then could have been used as a metric for evaluation.

Conclusion and Future Work

While there is room for improvement in this system, Emojinator was successful in meeting the objectives of generating emojis to represent emotions, making decisions on its own through AI and also being creative to some extent. Considering that not a lot of computational creativity work focuses on emoji generation, this system makes a unique contribution to existing research in this area. By incorporating soft computing techniques such as fuzzy logic, the system is also tolerant of approximations and imprecisions.

Overall, the emojis generated show potential for being used in computer-mediated communication. However, some parts of the software can be improved upon to make Emojinator more creative as a system.

In the future, the system can be further improved upon by making it more appreciative. This can be achieved by giving the system feedback on the emojis generated. This has also been done in previous works on emoji generation (Cunha et al. 2019) by making use of interactive evolutionary algorithms. This will also make the system more appreciative since it will know what the user likes and what the user does not like. This could result in the software becoming more creative, in line with the creative tripod approach.

Currently, the system also makes use of rules which define what kind of features it should have. This can be improved upon by training the system to identify on its own what kind of features an emoji should have. Actual human expressions can also be used to further enhance the system's understanding of emotions so that they reflect emotions better. The result of this would be greater autonomy for the system.

Research has also shown that a link exists between how personality, age and gender impact how emojis are interpreted. This is an interesting area to explore, and future user studies can also try to understand the relationship between emojis generated and the above-mentioned factors.

Link The source code for this system, along with the user study results and emojis generated can be found here: https://github.com/marziabil/emojis.

Acknowledgments

This work was undertaken by Marzia Bilwani in fulfilment of an MSc Computer Science project and dissertation, and was supervised by Anna Jordanous, whose support and guidance proved invaluable throughout.

References

Abbing, R. R.; Pierrot, P.; and Snelting, F. 2017. Modifying the universal. *Executing Practices* 17.

Boutet, I.; LeBlanc, M.; Chamberland, J. A.; and Collin, C. A. 2021. Emojis influence emotional communication, social attributions, and information processing. *Computers in Human Behavior* 119.

Churches, O.; Nicholls, M.; Thiessen, M.; Kohler, M.; and Keage, H. 2014. Emoticons in mind: An event-related potential study. *Social Neuroscience* 9.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. Technical Report SS-08-03.

Crystal, D. 2006. *Language and the Internet*. Cambridge, UK: Cambridge University Press.

Cunha, J. M.; Lourenço, N.; Correia, J.; Martins, P.; and Machado, P. 2019. Emojinating: Evolving emoji blends. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11453 LNCS.

Cunha, J. M.; Lourenço, N.; Martins, P.; and Machado, P. 2020. Visual blending for concept representation: A case study on emoji generation. *New Generation Computing* 38.

Cunha, J. M.; Martins, P.; and Machado, P. 2018a. Emojinating: representing concepts using emoji. In *Workshop Proceedings of the 26thth International inproceedings on Case-Based Reasoning*.

Cunha, J. M.; Martins, P.; and Machado, P. 2018b. How shell and horn make a unicorn: Experimenting with visual blending in emoji. In *Proceedings of the 9th International inproceedings on Computational Creativity, ICCC 2018*.

Dimson, T. 2015. Emojineering part 1: Machine learning for emoji trends. https://instagram-engineering.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine-learning-for-emoji-trends-7f5f9cb979ad.wufm6luuc.

Doble, A. 2015. Uk's fastest growing language is... emoji. https://www.bbc.com/news/newsbeat-32793732.

Ekman, P. 1999. *Basic Emotions*. John Wiley and Sons. 45–60.

EmojiList. 2022. Full emoji list, v14.0. https://unicode.org/emoji/charts/full-emoji-list.html.

EmojiMoji. 2022. Emojimoji. https://emblemmatic.org/emojimoji.

Emojipedia. 2022. Emojipedia. https://emojipedia.org/.

Fauconnier, G., and Turner, M. 2002. The way we think: conceptual blending and the mind's hidden complexities. *Choice Reviews Online* 40.

Fricker, R. D., and Schonlau, M. 2002. Advantages and disadvantages of internet research surveys: Evidence from the literature. *Field Methods* 14:347–367.

Guo, Z. X., and Wong, W. K. 2013. *Fundamentals of artificial intelligence techniques for apparel management applications*.

Herring, S. C., and Dainas, A. R. 2020. Gender and age influences on interpretation of emoji functions. *ACM Transactions on Social Computing* 3.

Jaeger, S. R.; Xia, Y.; Lee, P. Y.; Hunter, D. C.; Beresford, M. K.; and Ares, G. 2018. Emoji questionnaires can be used with a range of population segments: Findings relating to age, gender and frequency of emoji/emoticon use. *Food Quality and Preference* 68.

Joao Correia, Tiago Martins, P. M., and Machado, P. 2016. X-faces: The exploit is out there. In *Proceedings of the 7th International inproceedings on Computational Creativity*.

Koch, T. K.; Romero, P.; and Stachl, C. 2022. Age and gender in language, emoji, and emoticon usage in instant messages. *Computers in Human Behavior* 126.

Liapis, A. 2018. Recomposing the pokémon color palette. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10784 LNCS.

mathworks.com. 2022. Mamdani and sugeno fuzzy inference systems. https://www.mathworks.com/help/fuzzy/types-of-fuzzy-inference-systems.html.

Miller, H.; Kluver, D.; Thebault-Spieker, J.; Terveen, L.; and Hecht, B. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Proceedings of the 11th International inproceedings on Web and Social Media, ICWSM 2017*.

Paltoglou, G., and Thelwall, M. 2013. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing* 4.

Pavalanathan, U., and Eisenstein, J. 2015. Emoticons vs. emojis on twitter: A causal inference approach. http://arxiv.org/abs/1510.08480.

Phelps, E. A. 2006. Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology* 57.

Puyat, M. 2017. Emotigan: Emoji art using generative adversarial networks.

py5coding. 2022. Py5. http://py5coding.org/.

Radpour, N. D., and Bheda, V. 2017. Conditional generative adversarial networks for emoji synthesis with word embedding manipulation.

Rodríguez, L.-F.; Ramos, F.; and Wang, Y. 2012. Cognitive computational models of emotions and affective behaviors. *International Journal of Software Science and Computational Intelligence* 4.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39.

Scherer, K. R. 2005. What are emotions? and how can they be measured? *Social Science Information* 44.

Spolaor, S.; Fuchs, C.; Cazzaniga, P.; Kaymak, U.; Besozzi, D.; and Nobile, M. S. 2020. Simpful: A user-friendly python library for fuzzy logic. *International Journal of Computational Intelligence Systems* 13.

Stewart, L. H.; Ajina, S.; Getov, S.; Bahrami, B.; Todorov, A.; and Rees, G. 2012. Unconscious evaluation of faces on social dimensions. *Journal of Experimental Psychology: General* 141.

textfixer.com. 2022. The random choice generator tool. https://www.textfixer.com/tools/random-choice.php.

Walther, J. B., and D'Addario, K. P. 2001. The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review* 19.

Wang, Z.; Mao, H.; Li, Y. J.; and Liu, F. 2017. Smile big or not? effects of smile intensity on perceptions of warmth and competence. *Journal of Consumer Research* 43.

Wang, Y. 2007a. On the cognitive processes of human perception with emotions, motivations, and attitudes. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 1.

Wang, Y. 2007b. The theoretical framework of cognitive informatics. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 1.

Yannakakis, G. N., and Martinez, H. P. 2015. Ratings are overrated! *Frontiers in ICT* 2.

Yannakakis, G. N.; Cowie, R.; and Busso, C. 2018. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing* 12(1):16–35.

Yuasa, M.; Saito, K.; and Mukawa, N. 2011. Brain activity when reading sentences and emoticons: An fmri study of verbal and nonverbal communication. *Electronics and Communications in Japan* 94.

Zadeh, L. A. 1996. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems* 4.

Exploring Human Models of Innovation for Generative AI

Gualtiero B. Colombo, Hantao Liu, Roger M. Whitaker
School of Computer Science and Informatics
Cardiff University
Cardiff, CF24 4AG UK
ColomboG@cardiff.ac.uk, LiuH35@cardiff.ac.uk, whitakerrm@cardiff.ac.uk

Abstract

The ability to innovate is a precious commodity that humans are well-disposed to accomplishing. Currently, in the quest for development of artificial intelligence that is generative, blueprints for innovation are important for consideration. However those following principles of human innovation have been somewhat overlooked. The field of cultural evolution presents interesting models for explaining human innovation as evolutionary processes driven by intelligent biases. These offer approaches that can be followed in an algorithmic form by the machine, with particular degrees of freedom concerning bias. In this paper we take a first step in exploring cultural evolution models for generative AI. In particular, we use the concepts of cultural selection and biased transformation, where changes are driven by the bias imparted at different points of human decision making processes, in addition to social learning from others. We develop these approaches using a population of neural networks, each capable of drawing an image. We explore how neural networks and the resultant art evolve under these alternative interpretations of cultural evolution, using biases based on preferred prior images. The investigation suggests that bias must also evolve for innovation to persist, something which is given little emphasis in the literature.

Introduction

Interest in enabling computers to support creativity is something that is as old as computer science itself. Indeed, in an essay dating back to 1948 (Turing 1948), Alan Turing, a father of modern computing, speculated that different forms of search would need to support "intelligent machinery", and highlighted what he called "cultural search" as a process aligned to the mission of human creativity (Turing 1950). In very recent times we are now seeing considerable progress in artificial intelligence (AI) that is creative in open-ended scenarios, particularly through *generative AI*. Applications such as ChatGPT (Schulman et al. 2022), DALL-E 2 (Ramesh et al. 2022), and Stable Diffusion (Rombach et al. 2022) can create meaningful and complex content in response to limited instructions, for scenarios of complexity beyond which we have previously seen (Oppenlaender 2022). This is particularly the case for generative art and text, where the underlying models are highly dependent on scale (Galanter 2016; Boden and Edmonds 2009). For example, the training of large language models requires billions of words, bringing into question whether as an alternative, there are useful general principles through which meaningful creativity can established by alternative computational processes (Floridi and Chiriatti 2020; Dale 2021; Berns et al. 2021). To this end, current successful approaches include generative adversarial networks, where creativity is driven by competition (Tan et al. 2017), racheting training between a generating and discriminating neural network that results in high quality generation for open-ended problems including art (Shahriar 2022).

Less well considered are the processes underpinning the evolution of human innovation. The rationale for these models is strong, as humans have been able to innovate beyond all other species. Fundamental insights on the nature of creative processes have spanned both computing and psychology, with contributions such as those from Boden and Sawyer (Boden 2005; 2004; Sawyer 2011) indicating methods through which the individual mind can achieve creativity. These contributions serve to counter the illusion that creativity is a form of "magic", instead being methods that allow large search spaces to be navigated. It is also worth noting the often highly social nature of creativity, in that it is rarely achieved in isolation, and is progressively developed by building on the achievements of others, or from "*standing on the shoulders of giants*". This is something that Turing articulated in his early treatment of this subject (Turing 1948).

Today this area is recognised as *cultural evolution* (Tomasello 2009; Boyd and Richerson 1988; Mesoudi 2011), a cross-disciplinary endeavour that broadly seeks to understand how human innovations take hold (Tomasello, Kruger, and Ratner 1993). Here innovation has a specific meaning, representing the combination of *invention* and *social learning* (Paulus and Dzindolet 2008). In this context creativity is the process supporting invention, and this may influenced by others. Through cultural evolution, *cumulative culture* (Mesoudi and Thornton 2018) is now seen as a front runner in explaining how humans have become supreme innovators as compared to all other species. The main premise of cumulative cultural evolution is the concept of *ratcheting*, where improvements and novelty build without reverting to previous states over the longer term (Tennie, Call, and Tomasello 2009). This allows creativity to build in sophistication. It is argued that this presents simi-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

larities to Darwinian evolution (Mesoudi 2011), albeit with different mechanisms that allow change to happen much more quickly. Although there is significant debate about how cumulative culture results, at least two key models have emerged, known as *biased transformation* and *cultural selection* (Mesoudi 2021). These models expose the crucial role that human bias plays in executing cultural evolution, and can be approximated as simple algorithms that offer degrees of freedom as to their interpretation, configuration and sophistication. They can also function without necessarily pre-training, using association with memories to impart biases, which is often aligned with human decision making.

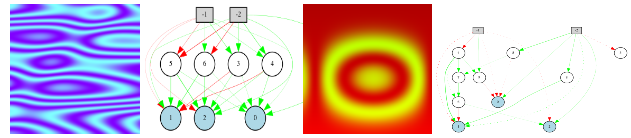
Contribution

Our overall interest is in techniques that are able to *persistently create artifacts that increasingly embody innovation and novelty*. Based on the success of humans in achieving this, in this paper we focus on two key mechanisms of cultural evolution, namely *cultural selection* and *biased transformation*. We adopt them as an inspiration for computational techniques where a small group of retained preferences are collectively engaged in creating new art represented through neural networks. Note that art is chosen as a vehicle to study creativity because it is readily accessible for human interpretation of innovation.

In general, innovation can be a challenging concept to measure because features that are innovative may not be foreseen, impeding quantitative measures. We use a general approach where instances of art are each drawn by a neural network and a set (or population) of images are maintained. Bias through images (represented by neural networks) are used to effectively represent preferences and a persistent memory, referred to as *preferred priors*. We explore ways in which bias can be applied through neural networks that are engaged in creating art, and we include the use of techniques from neuro-evolution to combine and impart bias. This approach allows directed modifications to be explored, where new images are created that are then available to repeatedly build upon in future. Since images are easily human interpretable we are able to gain a first-hand qualitative understanding of the role of bias in creative computational processes aligned to cultural evolution. This allows us to develop new insights and hypotheses.

The Neural Network Artist

Artificial neural networks (ANNs) are a mainstay of current AI, being robust to scaling and applicable across wide ranging scenarios. We focus on a novel form of ANNs called Compositional Pattern Producing Networks (CPPNs), that can be used to create images (Stanley 2007). Interesting images can be produced even from simple CPPN structures - an example is shown in Figure 1. CPPNs function by taking the x and y coordinate of an image's pixel as the input, and return as output the colour of the pixel. Thus each CPPN can be thought of as an artist that has created a painting. The structure of the CPPN and the activation functions used on the CPPN's nodes determine the form of the image that is produced. The previous use of CPPNs for image generation has been highly successful, for example being used in participatory crowdsourcing experiments (Secretan et al. 2008) that transformed abstract art to meaningful or interesting compositions. Using CPPNs provides a way to expose the "brain" of the virtual artist, and a novel point of influence. For example it is possible to make small mutations to a CPPN, such as a random change to the function of a node or the strength of connection between nodes (Stanley 2007), resulting in perturbations to the associated image. More directed changes can be invoked through combining CPPNs so that characteristics from one image can be used to influence another. This is non-trivial because the structure of CPPNs may vary making it challenging to map between such neural networks. However techniques from neuro-evolution can be used to create this integration, such as the crossover mechanism used in the NEAT algorithm (Stanley and Miikkulainen 2002; Stanley, D'Ambrosio, and Gauci 2009). Consequently, CPPNs and the approach to combining them through crossover give a means to impart bias in a cultural evolutionary process. These capabilities allow new computational models based on cultural evolution to be explored. In particular, it is interesting to see the extent to which they can be harnessed to generate novel artifacts with persistent creativity and increased complexity. This aspect gives an important insight concerning open-endedness, which is the grand-challenge of creating an algorithm that persistently creates innovations (Stanley 2019; Lehman, Stanley, and others 2008; Stanley, Lehman, and Soros 2017).



random CPPN Image e CPPN

Figure 1: CPPNs representing a simple randomly generated image and a more complex 'prior image'.

Experimental approach

To explore models of cultural evolution in this context, we address two directions for experimentation. Firstly, to understand the search space and the impact of bias, we consider biased and non-biased navigation through the search space, and the differences between them (Experiment 1). This requires consideration of the effects on changes to evolving a selected starting image. Secondly, to understand cultural models, we consider how approximations to biased transformation and cultural selection perform (Experiment 2), using retained memories of preferred prior images as the basis for bias. Note that necessarily, the evaluation here is subjective and exploratory, as a necessary first step to be further developed. The benefits of using art are in the human interpretability of the artifacts and their novelty, but this can only be judged by selective qualitative means. Nevertheless, this provides a useful basis to develop observations and hypotheses for more rigorous future investigation.

Experiment 1: Mutating a Single Image

What happens when we mutate a CPPN, or more specifically, how is the resulting image disrupted? This is a fundamental question underlying models of cultural evolution. Changes to artifacts, whether intentional or random, are the basis for *variation*. Variation underpins all evolutionary processes, without which artifacts remain in stasis. Variation opens up choice for further modification, and it is the basis for further random or directed change. The choices and modifications that are made determine how creativity emerges. While the human can impart intuition in selection and transformation decisions, using embedded skills, experiences and memories, the computer requires explicit bias to be programmed.

Algorithmic Approach

To assess this we compare two alternative strategies. Firstly taking a single starting point CPPN, denoted I, that has been randomly generated (and thus maps to a random image), we apply small random changes to the edge weighting or the activation function within a node from a hidden layer of the CPPN. This is successively repeated to the CPPNs that result, being equivalent to a random walk through the search space, representing an unguided and uninfluenced path. The general approach used is presented in Algorithm 1. This involves starting with a CPPN I (line 2) and evolving this using a simple procedure where n_s alternative random mutations are made to I (called the candidate set - line 7) from which one is randomly selected to update I (lines 8 and 9). This approach to randomisation is used so that it exploits a common framework of code. The output from Algorithm 1 gives a baseline to understand characteristics of a non-biased random approach.

Secondly, we consider biased transformation, where directed search is applied (Algorithm 2) in place of the small random changes considered in Algorithm 1. Algorithm 2 has two variations embedded within it, based on the strength of directed transformation. Initially a set of preferred prior images, denoted P, are defined (line 1). These provide the source of bias - in effect memories of interesting images represented by CPPNs. The starting CPPN for subsequent evolution is initalised in line 2. Note that this could also be one of the preferred priors.

The first step is to create a set of alternatives from I, which is called the *candidate set*. This is populated in two alternative ways (line 7), either mutations of I (called $biasedTrans_1$) or a crossover and mutation between I and a random CPPN from the set of priors (called $biasedTrans_2$). Crossover is a mechanism that combines two CPPNs and imparts characteristics from both to form a new CPPN, akin to creating an offspring. It is a non-trivial operation because CPPNs are not guaranteed to have the same structure. Here we use the crossover operator defined in (Stanley and Miikkulainen 2002). The resulting CPPN I' is subject to a small mutation (line 8) to guard against I or a prior being over-represented in the candidate set. Finally, a selection is made from the candidate set and this involves bias aligned to a particular prior. To achieve this, at each iteration, a preferred prior of interest is randomly chosen from P, denoted P_i (line 5). Then a subset of the candidate set is identified, involving a selection of most similar n_i images to P_i. This subset is called the *individuals set* (line 10). This excludes those images that are less well-related to P_i. Note that similarity is applied to the images rather than the CPPNs from which they are defined. This requires techniques from image processing and we employ a deep residual neural network Resnet (He et al. 2016), trained on the over 20 million images from the Imagenet dataset, further fine tuned with the methodology presented in (Wang et al. 2014). This allows general abstract features of images to drive similarity, such as shapes and patterns. It is ideal for our needs because it ensures similarity isn't based on over-fitting, being in more keeping with human intuition rather than precision. From the individuals set, then finally a random selection is made (line 11), which is updates I.

For exploratory purposes, a candidate set with cardinality 50 has been used, alongside an individuals set of size 10 and we run the algorithms for 50 iterations. Mutation settings for the CPPNs are set as equal to the default from (McIntyre et al. 2015).

Algorithm 1 Random Walk

1: **procedure** RANDOM WALK(Evolving CPPN I, Candidate Set Size n_s, Number of Iterations n^i)
2: $I \leftarrow P_j \in P$ **or** $I \leftarrow RandomImage$; $i = 0$
3: **while** $i < n^i$ **do**
4: CandidateSet = $\emptyset$; $j = 0$
5: **for** $j < n_s$ **do**
6: $I' = mutate(I)$
7: CandidateSet $\leftarrow$ CandidateSet $\cup\ I'$; j++;
8: $V' \leftarrow$ Select Randomly from CandidateSet
9: $I \leftarrow V'$; i++;

Algorithm 2 $biasedTrans_1$ and $biasedTrans_2$

1: **procedure** BIASEDTRANSFORMATION(Set of Prior Images P, Evolving CPPN I, CandidateSetSize n_s, IndividualsSetSize n_i, NumberOfIterations n^i)
2: $I \leftarrow P_j \in P$ **or** $I \leftarrow RandomImage$; $i = 0$
3: **while** $i < n^i$ **do**
4: CandidateSet = $\emptyset$, IndividualsSet = $\emptyset$; $j = 0$
5: Set Current Prior as $P_i \in P$
6: **for** $j < n_s$ **do**
7: $I' \leftarrow I$ **or** $\triangleright\ biasedTrans_1$
$I' \leftarrow$ Crossover(I, P_i) $\triangleright\ biasedTrans_2$
8: $I'' = mutate(I')$
9: CandidateSet $\leftarrow$ CandidateSet $\cup\ I''$; j++;
10: Select IndividualsSet $\subset$ CandidateSet as the Set of n_i Images in CandidateSet Most Similar to Current Prior P_i
11: $V' \leftarrow$ Select Randomly from IndividualsSet
12: $I \leftarrow V'$; i++;

Experiment 1: Results

Firstly we consider the effects of random walk through the search space, based on Algorithm 1. To demonstrate

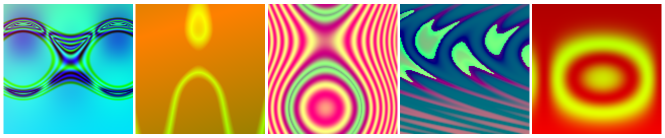


Image a Image b Image c Image d Image e

Figure 2: Example of five images represented through CPPNs that are used as candidates preferred priors and/or starting points in various experiments. These are chosen for having little similarity between them based on ResNet assessment.

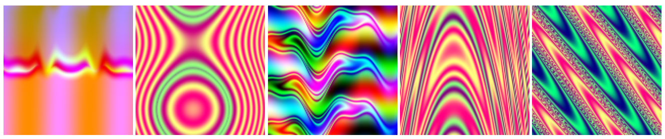


Image f Image g Image h Image i Image j

Figure 3: Example of five images represented through CPPNs that are used as candidates preferred priors in various experiments. These are chosen for having greater similarity between them based on ResNet assessment.

this, five sample images from Figure 2 and their associated CPPNs were taken as starting points, and the random evolution of resulting images were observed. Figure 4 shows interesting snapshots from the resulting sequence of images produced. Relatively quick changes occur across the images, with particularly interesting iterations highlighted in Figure 4, where the relationships between images can be observed. As expected, there are no overall patterns that can observed. Some random paths become complex in different ways (e.g., images a and d) while others remain similarly complex (e.g., image b), although they all tend to become more abstract and less well aligned to particular shapes that humans can identify with. Others drift to lower complexity (e.g., c and e). What is evident though is the rich variety of images that can be readily generated and how easy it is to transverse the search space.

In contrast to a random walk, the results from biased transformation (Algorithm 2) are stark, where the impact of biased transformation is significant. Figures 5 and 6 demonstrate this using a common starting CPPN (see Figures 1 and 2), and they present selections from some of the most interesting images created across different runs with alternative random seeds. These are of course subjective selections, but are representative of the search space. In each of Figures 5, 6, 7 and 8 different priors are applied, ranging from two to five priors. Also two alternative starting images are used, to provide a further comparison. Finally we also consider difference in applying a set of highly similar priors (Figure 3) as compared to a set that are mutually dissimilar, based on ResNet similarity (Figure 9).

Using qualitative inspection from experimentation, we draw the following observations and hypotheses. Firstly, $biasedTrans_2$ has a much stronger evolutionary impact in terms of diversity of interesting images than for $biasedTrans_1$. In other words, introducing crossover provides a strong influence to direct the evolution towards creative areas of the search space related to preferred priors. Secondly, although a minimal number of preferred priors will support the discovery of interesting solutions, additional priors appear to increase the diversity and complexity of the most interesting images that are discovered. This is especially the case when crossover is involved ($biasedTrans_2$). However, fully displaying this using a limited number of images is challenging. Thirdly, in all cases, using alternative random seeds is sufficient to drive the evolution in diverse directions across the search space. This seemed to be amplified when a greater number of priors were involved, or when crossover was employed ($biasedTrans_2$). Fourth, it is evident that both similar and dissimilar sets of prior images can promote creativity, while the dissimilar priors seem to influence shape formation, while similar priors seem to add more intricate detail to images. Finally, the results seem to affirm that the concept of similarity to retained memories (or preferred priors) is sufficient to drive creativity within a few iterations without new images being directly tied to the priors from which they have been influenced.

Experiment 2: Cultural Evolution on a Population of Images

To further explore models of cultural evolution, it is necessary to introduce a population of alternative artifacts that are available to be updated, providing a diversity through which innovations can accumulate and transfer between individuals. This is a basis for two key models of cultural evolution, namely *cultural selection* and *biased transformation* (Mesoudi 2021). These models allow external influence from multiple sources each represented by a *preferred prior* image to possibly influence decision making when acting upon a population of artifacts of some description. The difference between cultural selection and biased transformation concerns where bias comes into play - either at the point of selection of artifact to modify (cultural selection) or the way in which they are modified (biased transformation). In reality both these elements may combine (Mesoudi 2021) but it is prudent, from an exploratory perspective, to understand the differences between alternative models.

Algorithmic Approach

We use the algorithmic framework analogous to that described in the previous section and focus on evolving a population of random CPPNs aligned to different models of cultural evolution. Firstly we focus on the effects of bias at the selection stage, aligned to *cultural selection*. In Algorithm 3 we evolve a population of N CPPNs (set Pop) by replacing at each generation a subset of Pop, denoted *ReplacedSet*, and replacing it with another set of CPPNs, denoted *NewIndSet*.

To decide how *NewIndSet* is composed, we firstly create a set, *CandidateSet* (lines 8-11) from the population. This can be created in two alternative ways (line 9), either through mutating a randomly selected individual (which we

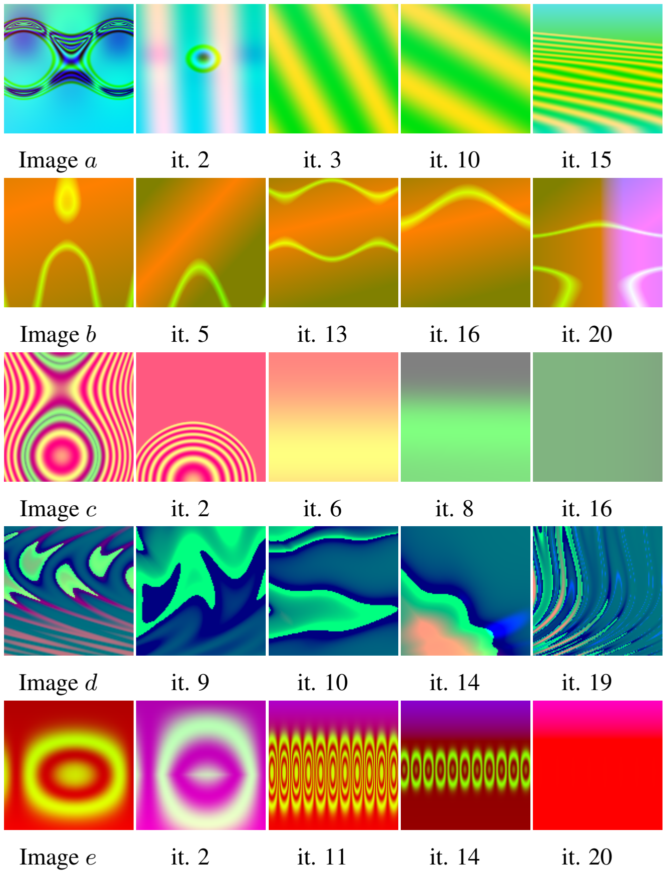


Figure 4: Examples of representative and interesting images produced under Algorithm 1 (random walk). Each row represents a sample of images selected from random evolution from five starting points (Images a - e). Numbers under images indicate the iteration from which they are taken.

call *random selection*), or by making the selection based on weighted similarity with a randomly selected prior P_i (line 6), representing a form of direct bias (Boyd and Richerson 1988), and then applying a mutation. We call this approach *cultural selection*. The *NewIndSet* is then randomly selected as a subset of *CandidateSet*. Finally, the set of images that are least similar to P_i are then removed from the population *Pop* (*ReplacedSet* in lines 16-18). This biased removal is important in various approaches to cultural evolution (Boyd and Richerson 1988; Dawkins and others 1996).

Algorithm 4 evolves a population of CPPNs based on *biased transformation*, where the transformation of artifacts from the population are subject to biases. This contrasts to Algorithm 3 where the bias is applied to selection. As in Algorithm 3, at each generation Algorithm 4 takes a subset of Pop, denoted *ReplacedSet*, and replaces it with another set of CPPNs, denoted *NewIndSet*. Members of the *CandidateSet* are initially selected at random (line 9). These are then either directly mutated under $biasedTransformation_1$ or crossover occurs with P_i, representing a form of guided variation (Boyd and Richerson 1988), denoted $biasedTransformation_2$ (lines 10-

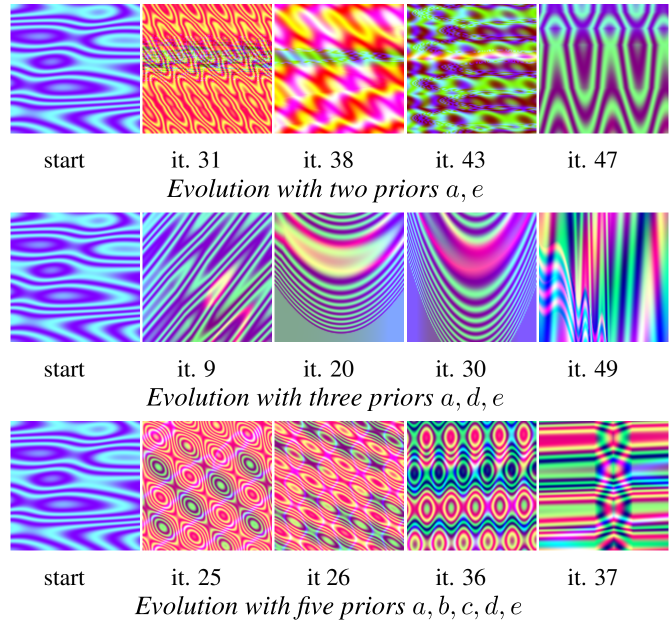


Figure 5: Examples of the most creative and interesting images produced under $biasedTrans1$ (Algorithm 2) for two (top), three (middle), and five (bottom) prior images using a randomly generated image as starting point. Numbers under images indicate the iteration from which they are taken.

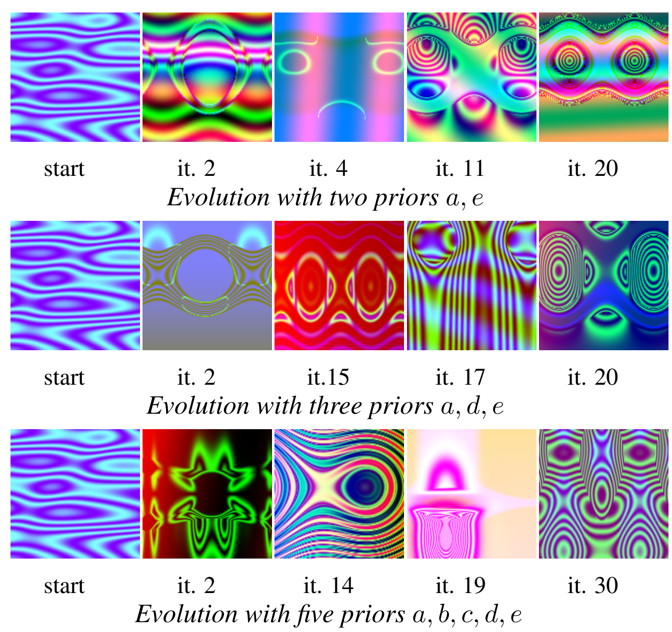


Figure 6: Examples of the most creative and interesting images produced under $biasedTrans2$ (Algorithm 2) for two (top), three (middle), and five (bottom) prior images using a randomly generated image as starting point. Numbers under images indicate the iteration from which they are taken.

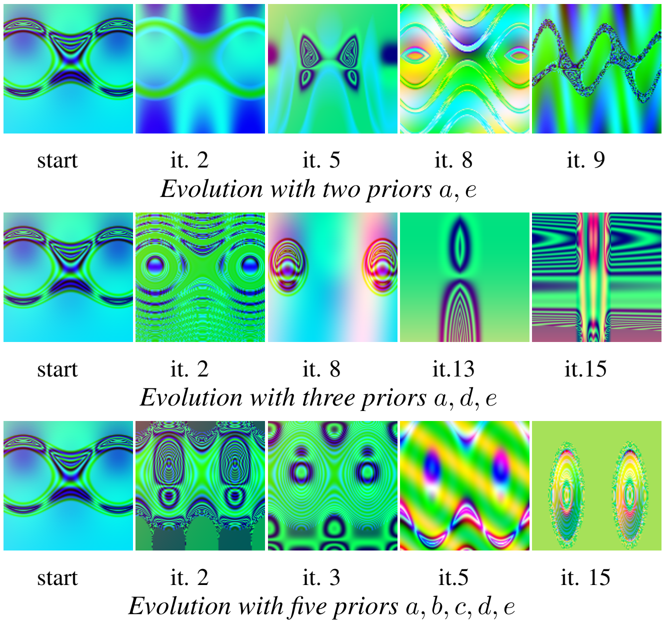


Figure 7: Examples of the most creative and interesting images produced under $biasedTrans1$ (Algorithm 2) for two (top), three (middle), and five (bottom) prior images using Image a as starting point. Numbers under images indicate the iteration from which they are taken.

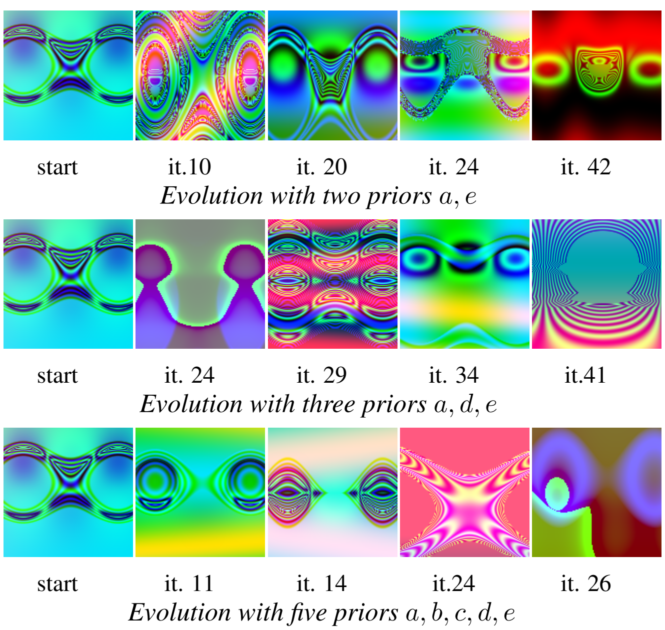


Figure 8: Examples of the most creative and interesting images produced under $biasedTrans2$ (Algorithm 2) for two (top), three (middle), and five (bottom) prior images using Image a as starting point. Numbers under images indicate the iteration from which they are taken.

11). From *CandidateSet*, we then select a subset *NewIndSet* that is most similar to P_i (lines 13-15). Finally, the set of

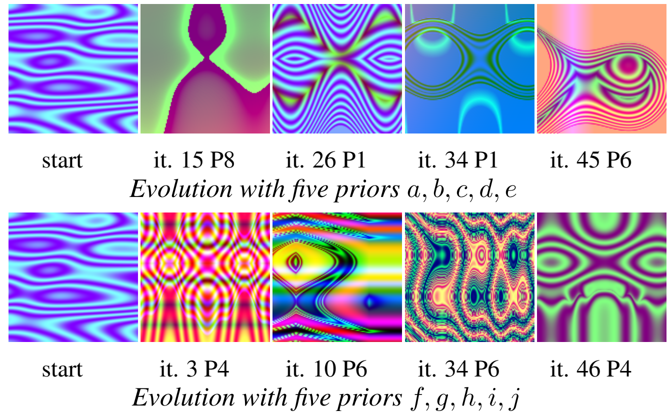


Figure 9: Examples of the most creative and interesting images produced under $biasedTrans2$ (Algorithm 2) for similar (top) and dissimilar (bottom) sets of priors using a randomly generated image as starting point. Numbers under images indicate the iteration from which they are taken.

images that are least similar to P_i are then removed from *Pop* (*ReplacedSet* in lines 16-18) and *NewIndSet* is added to *Pop*. This replacement approach follows that in Algorithm 3. For exploratory purposes, a candidate set with cardinality 250 has been used, alongside an individuals set of size 10, a replaced set also of size 10, and we run the algorithms for 50 generations.

Experiment 2: Results

Firstly we consider evolving a population of CPPNs ($N = 50$) and adopt only one preferred prior, taking image e from Figure 3. This is designed to test the potential convergence differences between the variations in Algorithms 3 and 4. The cultural evolution literature (Mesoudi 2011; Boyd and Richerson 1988) indicates that while both culture selection and biased transformation support convergence in simple fitness-based models, biased transformation is quicker.

The results from Figure 10 show that biased transformation has a significant effect in directing the evolution based on the preferred prior, as compared to both cultural and random selection. This is particularly the case for $biasedTransformation_2$. Similarity here is measured using ResNet-based measure (Wang et al. 2014). Note that all approaches involve biased removal of CCPNs from the population at the point at which replacements are added. The results also show that there are fundamental limitations in applying only a single point of bias, in this case a single preferred prior, because evolution gets drawn towards the original point of bias, as seen in Figure 11. This is also noted in cultural evolution treatments of biased transformation.

Secondly, based on using a set of five preferred priors as bias, we consider the characteristics of cultural evolution based on Algorithms 3 and 4. In Figure 14 we present a selection of images representative of the most novel and creative images from across different generations for each of the four techniques presented through Algorithms 3 and 4.

Algorithm 3 Random Selection and Cultural Selection

procedure RANDOMSELECTION / CULTURAL SELECTION (NumberofGens G, EvolvingPupulationOfImages Pop, SetofPriorImages P, PopSize N, CandidateSetSize n_s, NewIndSetSize n_i, ReplacedSetSize n_r)
- $NewIndSet$=$\emptyset$,$CandidateSet$=$\emptyset$,$ReplacedSet$=$\emptyset$
- Set Pop as a Population of Random Images; $g = 0$
- **while** g< G **do**
  - Select Randomly Current Prior P_i from P
  - $i = 0$; $j = 0$; $k = 0$
  - **for** $i < n_s$ **do**
    - $I \leftarrow$ Select Randomly from Pop ▷ rand. sel.
    - **or**
    - $I \leftarrow$ Select from Pop Proportionally to Similarity to P ▷ cult. sel.
    - $I' \leftarrow mutate(I)$
    - $CandidateSet \leftarrow CandidateSet \cup I'$; i++
  - **for** $j < n_i$ **do**
    - $V \leftarrow$ Select Randomly from CandidateSet
    - $NewIndSet \leftarrow NewIndSet \cup V$; j++
  - **for** $k < n_r$ **do**
    - $K \leftarrow$ Set of $ReplacedSetSize$ Images in Pop Least Similar to P_i
    - $ReplacedSet \leftarrow ReplacedSet \cup K$; k++
  - $Pop \leftarrow Pop - ReplacedSet$
  - $Pop \leftarrow Pop \cup NewIndSet$; g++

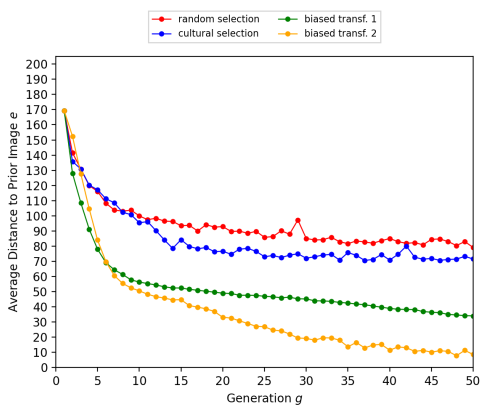
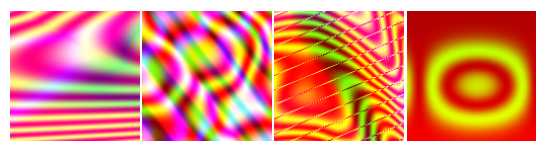


Figure 10: Average distance of population to image e as the preferred prior produced from the variations in Algorithms 3 and 4.

Algorithm 4 $biasedTrans_1$ and $biasedTrans_2$

procedure BIASEDTRANSFORMATION (NumberofGens G, EvolvingPupulationOfImages Pop, SetofPriorImages P, PopSize = N, CandidateSetSize n_s, NewIndSetSize n_i, ReplacedSetSize n_r)
- $NewIndSet$=$\emptyset$,$CandidateSet$=$\emptyset$,$ReplacedSet$=$\emptyset$
- Set Pop as Population of Random images; $g = 0$
- **while** g< G **do**
  - Select Randomly Current Prior P_i from P
  - $i = 0$; $j = 0$; $k = 0$
  - **for** $i < n_s$ **do**
    - $I \leftarrow$ Select Randomly from Pop
    - $I' \leftarrow I$ ▷ $biasedTrans_1$
    - **or**
    - $I' \leftarrow crossover(I, P_i)$ ▷ $biasedTrans_2$
    - $I''= mutate(I')$
    - $CandidateSet \leftarrow CandidateSet \cup I'$; i++
  - **for** $j < n_i$ **do**
    - $V \leftarrow$ Set of $NewIndSetSize$ Images in $CandidateSet$ Most Similar to P_i
    - $NewIndSet \leftarrow NewIndSet \cup V$; j++
  - **for** $k < n_r$ **do**
    - $K \leftarrow$ Set of $ReplacedSetSize$ Images in Pop Least Similar to P_i
    - $ReplacedSet \leftarrow ReplacedSet \cup K$; k++
  - $Pop \leftarrow Pop - ReplacedSet$
  - $Pop \leftarrow Pop \cup NewIndSet$; g++



random selection | cultural selection | biased trans. 1 | biased trans. 2

Figure 11: Images with greatest similarity to image e as the preferred prior produced from the variations of Algorithms 3 and 4 over 50 generations.

From inspection and multiple trials, greater levels of creativity seems apparent under biased transformation. Aligned to these experiments, we also track the average similarity in the population of CPPNs as compared to each of the five priors. This is presented for *cultural selection* (Figure 12) and $biasedTransformation_2$ (Figure 13). We note that there are considerable differences - while Figure 12 exhibits general trends towards similarity and convergence, the alternative is true in Figure 13.

Based on our qualitative observation from experimentation we draw the following observations and hypotheses. Firstly, although bias can drive creativity, it can also equally restrict creativity and innovation if biases are restricted, as seen when only a single prior is applied. Thus multiple sources of bias have an important effect on the creativity that is achieved. Secondly, multiple sources of bias, combined with the evolution of a population of images rather than evolution of a single image (Experiment 1) provides the opportunity for much diversity of images to emerge. This is more strongly felt under biased transformation, when the transformation is heavily directed through crossover-based techniques ($biasedTransformation_2$). Finally, from observing Figures 12 and 13 we hypothesise that tensions between biases can drive creativity and the innovation that results. In effect biases, when heavily directed through biased transformation, are steering a path through interesting elements of the search space, seemingly allowing more scope to evolve shapes for example, rather than only adding complexity to an existing form of image.

ISBN: 978-989-54160-5-9

Conclusion and Future Work

This study has brought together diverse techniques from neural networks, neuro-evolution and visual computing to support a new exploratory approach for harnessing computational creativity. These techniques have been used to explore whether fundamental human models of innovation, known as *cultural evolution*, can inspire new computational techniques to generate creativity from minimal user input and computational forms of bias. Models of cultural evolution represent important techniques because they capture the ways in which humans have been supremely successful as innovators. Art provides an excellent vehicle for exploring computational techniques in this context, with neural networks creating images for human interpretation of novelty. Our initial findings show prospects for new computational techniques based on cultural evolution. A key issue concerns the role of bias, and *biased transformation* in particular, shows promise. Imparting bias in a computational form can be a challenge, and the approach undertaken here supports the idea that retained memories, and abstract similarity to them, can function as an effective method. In other words, ideas from the past, or embedded preferences, can shape the creation of new artifacts in unforeseen and novel ways.

To understand the impact of biases we explored biased and non-biased navigation through the search space. Furthermore, we considered how approximations to biased transformation and cultural selection perform, to understand how these cultural models impact on algorithmic creation. Our investigation highlights the neural networks' ability to create images that evolve based on biased transformation, suggesting that further development should focus on the issue of bias and also that bias must evolve for innovation to persist. Novel forms of learning can be considered to allow adaption of bias in step with the level of complexity in the population. Little emphasis is given to the dynamics of bias in the cultural evolution literature, which often features snapshots of dynamic behaviour or static fitness functions as a proxy for bias. However machine learning and computational evolutionary techniques offer new prospects for achieving this. We believe that this could be an important aspect in developing persistent innovation aligned to open-endedness (Stanley, Lehman, and Soros 2017).

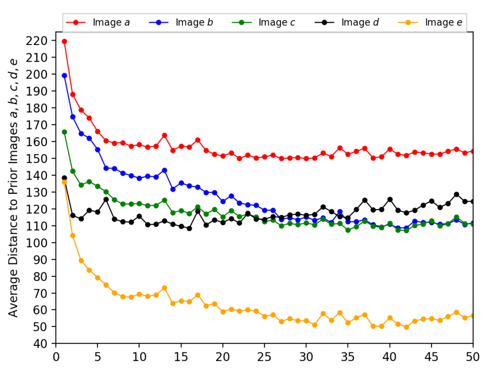


Figure 12: Average distance of population to the each of five prior images a, b, c, d, e over generations under *cultural selection* (Algortihm 3).

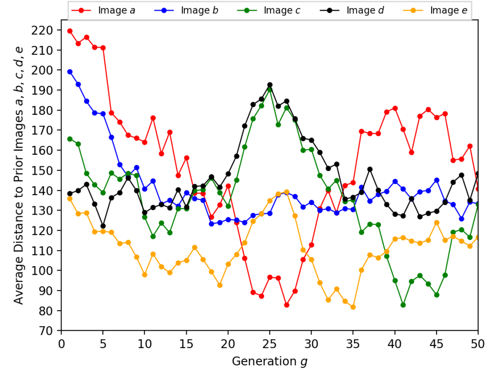


Figure 13: Average distance of population to each of the five prior images a, b, c, d, e over generations under $biasedTrans2$ (Algorithm 4).

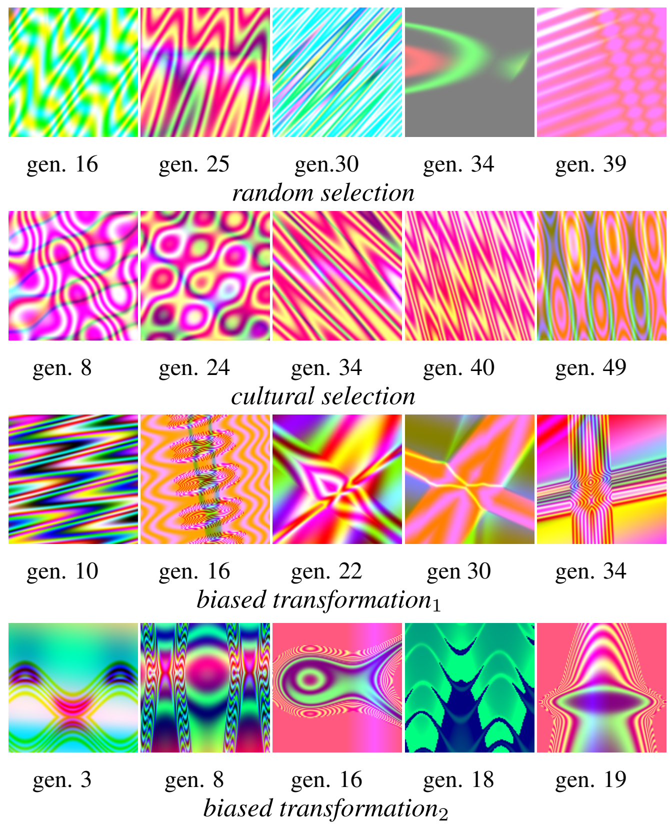


Figure 14: Examples of the most creative and interesting images from the population produced under the variations in Algorithms 3 and 4. A population of random images was used as starting point. Numbers under images indicate the generation from which they are taken.

Author Contributions

All authors contributed to the model, experimental design and analysis. Author 1 led the implementation. Author 2 provided additioanl expertise concerning visual computing. All authors contributed to drafting and reviewing the manuscript.

References

Berns, S.; Broad, T.; Guckelsberger, C.; and Colton, S. 2021. Automating generative deep learning for artistic purposes: Challenges and opportunities. *arXiv preprint arXiv:2107.01858*.

Boden, M. A., and Edmonds, E. A. 2009. What is generative art? *Digital Creativity* 20(1-2):21–46.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Boden, M. A. 2005. What is creativity? In *Creativity in human evolution and prehistory*. Routledge. 27–55.

Boyd, R., and Richerson, P. J. 1988. *Culture and the evolutionary process*. University of Chicago press.

Dale, R. 2021. Gpt-3: What's it good for? *Natural Language Engineering* 27(1):113–118.

Dawkins, R., et al. 1996. *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. WW Norton & Company.

Floridi, L., and Chiriatti, M. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30:681–694.

Galanter, P. 2016. Generative art theory. *A companion to digital art* 146–180.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Lehman, J.; Stanley, K. O.; et al. 2008. Exploiting open-endedness to solve problems through the search for novelty. In *ALIFE*, 329–336.

McIntyre, A.; Kallada, M.; Miguel, C. G.; Feher de Silva, C.; and Netto, M. L. 2015. neat-python.

Mesoudi, A., and Thornton, A. 2018. What is cumulative cultural evolution? *Proceedings of the Royal Society B* 285(1880):20180712.

Mesoudi, A. 2011. *Cultural evolution - How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. University of Chicago Press.

Mesoudi, A. 2021. Cultural selection and biased transformation: two dynamics of cultural evolution. *Philosophical Transactions of the Royal Society B* 376(1828):20200053.

Oppenlaender, J. 2022. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, 192–202.

Paulus, P. B., and Dzindolet, M. 2008. Social influence, creativity and innovation. *Social Influence* 3(4):228–247.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sawyer, R. K. 2011. *Explaining creativity: The science of human innovation*. Oxford university press.

Schulman, J.; Zoph, B.; Kim, C.; Hilton, J.; Menick, J.; Weng, J.; Uribe, J.; Fedus, L.; Metz, L.; Pokorny, M.; et al. 2022. Chatgpt: Optimizing language models for dialogue.

Secretan, J.; Beato, N.; D Ambrosio, D. B.; Rodriguez, A.; Campbell, A.; and Stanley, K. O. 2008. Picbreeder: evolving pictures collaboratively online. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1759–1768.

Shahriar, S. 2022. Gan computers generate arts? a survey on visual arts, music, and literary text generation using generative adversarial network. *Displays* 102237.

Stanley, K. O., and Miikkulainen, R. 2002. Evolving neural networks through augmenting topologies. *Evolutionary computation* 10(2):99–127.

Stanley, K. O.; D'Ambrosio, D. B.; and Gauci, J. 2009. A hypercube-based encoding for evolving large-scale neural networks. *Artificial life* 15(2):185–212.

Stanley, K. O.; Lehman, J.; and Soros, L. 2017. Open-endedness: The last grand challenge you've never heard of. *While open-endedness could be a force for discovering intelligence, it could also be a component of AI itself.*

Stanley, K. O. 2007. Compositional pattern producing networks: A novel abstraction of development. *Genetic programming and evolvable machines* 8(2):131–162.

Stanley, K. O. 2019. Why open-endedness matters. *Artificial life* 25(3):232–235.

Tan, W. R.; Chan, C. S.; Aguirre, H. E.; and Tanaka, K. 2017. Artgan: Artwork synthesis with conditional categorical gans. In *2017 IEEE International Conference on Image Processing (ICIP)*, 3760–3764. IEEE.

Tennie, C.; Call, J.; and Tomasello, M. 2009. Ratcheting up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1528):2405–2415.

Tomasello, M.; Kruger, A. C.; and Ratner, H. H. 1993. Cultural learning. *Behavioral and brain sciences* 16(3):495–511.

Tomasello, M. 2009. *The cultural origins of human cognition*. Harvard university press.

Turing, A. M. 1948. Intelligent machinery. Report for National Physical Laboratory. Reprinted in Ince, D. C. (editor). 1992. Mechanical Intelligence: Collected Works of A. M. Turing. Amsterdam: North Holland. Pages 107127. Also reprinted in Meltzer, B. and Michie, D. (editors). 1969. Machine Intelligence 5. Edinburgh: Edinburgh University Press.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1386–1393.

Towards the Automatic Evaluation of Visual Balance for Graphic Design Posters

Daniel Lopes, João Correia, Penousal Machado
University of Coimbra, CISUC, DEI and LASI, University of Minho
{dfl,jncor,machado}@dei.uc.pt

Abstract

Being able to evaluate aesthetics automatically is one of the fundamental needs for creating robust and autonomous computational creativity systems. In Graphic Design (GD), many aesthetic features might need to be considered simultaneously to properly evaluate GD artefacts, e.g. their visual relation to the concept of the work, legibility, innovation degree and the personal taste of the target public. Another relevant feature is the balance of the elements in the composition. This paper presents and tests an approach for evaluating the page balance of GD posters. Furthermore, it compares the evaluation computed by the developed method with the evaluation made manually by graphic designers and other creative practitioners. The results suggest the presented method can reasonably emulate the opinion of the human evaluators concerning the page balance of the presented posters. Moreover, for the presented setup, the results indicate a possible correlation between page balance and visual pleasantness, i.e. between the former and the personal taste of the human evaluators.

Introduction

More and more, Computational Creativity (CC) techniques have been explored to approach Graphic Design (GD) challenges, e.g. to speed up the GD creative process or aiding the exploration of innovative visual solutions.

To create CC systems that are capable of generating helpful GD solutions, one must first be able to create objective metrics to describe the quality of the expected designs. Nevertheless, due to the subjectivity of GD aesthetics, creating capable metrics for evaluating GD is still an open problem.

GD evaluation metrics may focus, for example, on the concept of the work, the legibility of the contents, innovation degree and other even more subjective features such as the personal taste of the target public. Another relevant feature is visual balance, which often relates to the visual weight of the items in the composition, on each side of a given axis.

Building on top of existing work (Harrington et al. 2004; Lok, Feiner, and Ngai 2004), this paper presents and tests a practical method to evaluate the page balance of GD posters. To do that, the brightness and position of each pixel in a given poster are considered to calculate a centre of mass (CM). The closer it is to a reference axis, the better the evaluation. Different axes and combinations of axes were tested. The evaluation values for each axis are weighed and summed up to calculate an overall evaluation value.

120 GD posters created by different graphic designers and gathered from a variety of sources, e.g. *typographic-posters.com*, *posters.calarts.edu* or websites from GD studios, were evaluated both automatically, using our method, and manually, by means of a user survey made with graphic designers and CC practitioners working on GD.

Subsequently, the automatically and manually obtained evaluation values were compared to assess whether or not the presented method could reasonably modulate human perception of page balance, at least, according to the opinion of the respondents of the conducted survey. Furthermore, studying the hypothesis of a correlation between page balance and visual pleasantness, the respondents were also asked how much they liked the respective posters.

The results suggest the proposed method could reasonably emulate the opinion of the respondents concerning the page balance of the presented posters. Moreover, the results indicate a possible correlation between page balance and visual pleasantness, at least, for the current experimental setup.

Related Work

As suggested before, this paper aims to contribute by introducing and testing a practical method to evaluate the visual balance of GD posters (2D), aiding the creation of CC systems for the generation of GD artefacts, such as posters. Hence, this section reviews existing work on the generation of 2D page layouts, especially focusing on page balance and CC systems for generating multipurpose GD layouts, i.e. which can be helpful in different GD briefings.

The automatic generation of multipurpose page layouts for GD has been done by numerous authors using different techniques. Constraint-based approaches are often used for displaying and aligning items on pages, e.g. using grid systems (Feiner 1988; Ferreira and others 2019; Cleveland 2010) or predefined templates (Jacobs et al. 2004). However, such systems are often unable to evaluate the generated results, so visual quality is usually controlled by humans or by restrictive hard-coded constraints, excluding such approaches from the CC domain.

Interactive Evolutionary Computation (IEC) has also been endorsed to generate GD layouts (Klein 2016; Kitamura and Kanoh 2011; Önduygu 2010). The shortcoming of IEC is the human users must still evaluate the generated results/candidates to drive the generation process.

One can also identify hybrid approaches in which the system automatically evolves layouts by fully filling in pages with a given number of text boxes, and the users are only asked to evaluate which they like the most (Rebelo et al. 2018). Nevertheless, such an approach to puzzle items into pages does not fit more generic contexts.

Geigel and Loui (2003) explored a more generic approach by constraining the layouts according to a number of hard-coded aesthetic metrics. Visual balance was automatically controlled by assessing page symmetry. However, such a metric can be reductive for defining visual balance, e.g. visual weight approaches (Harrington et al. 2004; Lok, Feiner, and Ngai 2004) can assess symmetry along with many other visual balance circumstances.

Automatic evolutionary computation (AEC) is one of the techniques that can benefit from automatic visual evaluation metrics, such as the one presented in this paper. AEC has demonstrated its potential to find solutions to complex problems (Stanley and Miikkulainen 2002), including in computational art contexts (Machado and Cardoso 2002) and some particular GD tasks, such as the generation of modular typography (Martins et al. 2016).

On the GD posters domain, there has been work using computer vision to automatically retrieve insights about whether the public is more or less interested in a given candidate poster and assigning fitness accordingly, i.e. the more a person looked at a poster, the better the fitness (Rebelo et al. 2017). While the latter work could only generate background variations on a single poster layout, there has also been work towards multipurpose systems. For example, to approximate existing layouts using different page items (Lopes, Correia, and Machado 2022).

Lastly, there has been work using Machine Learning (ML) techniques to learn features of existing layouts so one can generate new ones accordingly. For example, by learning how to drive the public's attention to given zones of the layout, detecting alignment or understanding hierarchical features (O'Donovan, Agarwala, and Hertzmann 2014), or generating layouts according to the semantic information of the page items (Zheng et al. 2019).

Specifically on the page balance domain, and besides symmetry (Geigel and Loui 2003), Harrington et al. (2004) proposed to assess visual balance by calculating the CM and measuring its distance to the centre of the page or, alternatively, assessing the difference between the visual weight on the left and on the right side of the page. The shortcoming is the authors' balance calculation is based on the average brightness of each page item, i.e. the metric is not well-fitted to be applied to raster images, especially if they're too complex. Furthermore, the average brightness of an item might be misleading if the object is visually heavier on one of its sides. Lok, Feiner, and Ngai (2004) used edge detection to assess the size, position and brightness of the page items and therefore calculated weight maps. A shortcoming of the latter approach is the page items were assumed to be uniformly weighted. As stated by the authors, a pixel-based approach might better reflect the way humans evaluate layouts.

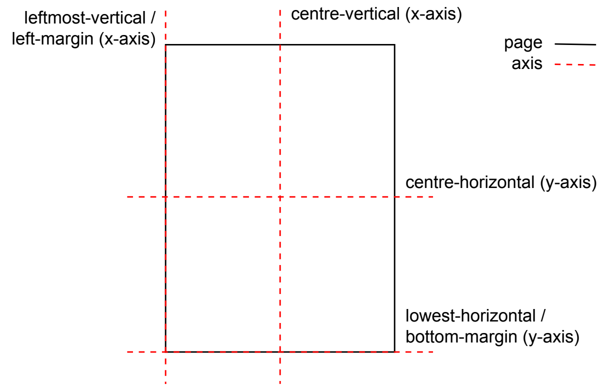


Figure 1: Page axes tested.

Approach

Inspired by the work of Harrington et al. (2004) and Lok, Feiner, and Ngai (2004), this paper presents a pixel-based method to evaluate visual balance in 2-dimensional GD artefacts, especially focusing on posters.

The proposed method, implemented in *JavaScript*, takes as input one PNG image of any size and ratio. However, to improve performance, we automatically resize the input image to 400 pixels wide. Height is set proportionally.

The CM of the given image is calculated considering the brightness and the position of each pixel. First, the default CM is assigned to a vector in the centre of the page (centred vertically and horizontally). Then, each pixel is assigned a weight equal to its inverted normalised brightness, i.e. 0 standing for brighter values and 1 standing for darker ones, so darker pixels were considered visually heavier, as often white pixels stand for emptiness/white page. This value is then squared, emphasising the differences between lighter and darker pixels. A vector referring to the position of the given pixel is then multiplied by its respective weight. Lastly, the resulting vector is added to the default CM vector. This way, darker pixels will more strongly attract the CM in their direction to the detriment of lighter ones.

After assessing the CM of the image, the distance to one or more axes can be calculated to estimate balance. In the following experiments, we evaluated posters considering the centre-vertical, centre-horizontal, leftmost-vertical (left-margin) and the lowest-horizontal (bottom-margin) axes (see Figure 1), either alone or mixing two axes together. The left-margin axis was tested in detriment to the right-margin one since the gathered posters communicate using left-to-right writing. To mix axes together, their distances to the CM were weighed and then summed up.

The full code can be downloaded from *GitHub* at *github.com/danifslopes/Visual-Balance-Evaluation*.

Experimental Setup and Analysis

As mentioned before, the system was tested by evaluating a dataset of 120 posters gathered from various sources. First, to more easily study the impact of the item's distribution on the page, we started by designing a set of 30 posters com-

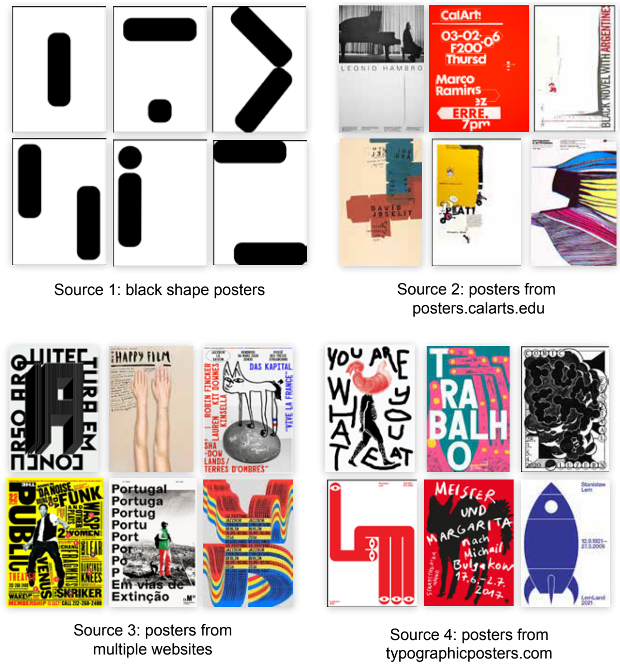


Figure 2: Examples of evaluated posters, grouped by type of source. (1) posters designed on purpose by our research team; (2) posters from the *posters.calarts.edu* archive; (3) posters from well-known GD studios, gathered from multiple websites; (4) posters from *typographicposters.com*.

posed of a maximum of two black geometric shapes over white background, positioned in varied dispositions. We refer to these as black shape posters. Another 30 posters were gathered online from *posters.calarts.edu*, a poster archive containing diverse posters quality-wise. The third set of 30 posters was gathered from several websites from well-known GD studios. Lastly, 30 posters were gathered from *typographicposters.com*, an online archive in which graphic designers worldwide upload their poster designs. Typically, this archive includes work by experienced designers as the users need to be approved by the administrators of the archive. Refer to Figure 2 for examples of posters from each of the aforementioned sources.

As the goal of proposed method is to aid the creation of AEC systems for GD, a group of 25 graphic designers and CC practitioners working on GD were asked to evaluate the posters concerning their visual balance and visual pleasantness. We refer to this as manual evaluation.

More specifically, all the respondents had a GD background, except for two who did not. Even so, these were working on CC for GD purposes. Also, all had Portuguese nationality except for one Brazilian living in Portugal for 2 years already. Except for 6 of them, all the respondents worked or studied at the University of Coimbra at the time the survey was conducted.

The respondents were asked, from 0 to 10, (i) *"How visually balanced do you think the posters are? Please, ignore whether you like them or not. Do not consider whether or not you like the colours, typefaces or other graphics"* and (ii) *"How aesthetically pleasing the posters seem to you, regardless of why?"*. Each respondent evaluated 24 posters, 6 of each type.

Secondly, the values resulting from manual evaluation were compared with the values resulting from automatic evaluation. To do that, the method was run over the 120 posters using different parameters. We tested evaluating the posters considering the following axes and combinations of axes: (i) Centre-Vertical alone (CV); (ii) Centre-Horizontal alone (CH); (iii) Left-Margin alone (L); (iv) Bottom-Margin alone (B); (v) Centre-Vertical & Centre-Horizontal (C+C); (vi) Left-Margin & Bottom-Margin (L+B); (vii) Centre-Vertical or Left-Margin whichever closer to CM & Centre-Horizontal (C/L+C); (viii) Centre-Vertical & Centre-Horizontal or Bottom-Margin whichever closer to CM (C+C/B); (ix) Centre-Vertical or Left-Margin whichever closer to CM & Centre-Horizontal or Bottom-Margin, also, whichever closer to CM (C/L+C/B).

For getting a unique balance value, whenever two axes were combined, their normalised distances to CM were weighted and summed up. For the vertical and horizontal axes respectively, we tested the following weights: [0.5, 0.5], [0.25, 0.75] and [0.75, 0.25].

To compare manual and automatic evaluation, we averaged the manual evaluation values and then calculated the (i) average distance (the closer to 0 the better) and (ii) the cosine similarity (the closer to 1 the better) between the manual and the automatic evaluation values. For a better comparison between metrics, the average distance was inverted, turning it into an average similarity value instead[1], i.e. the closer the value is to 1, the better.

Moreover, we tested similarities by (i) considering all 120 posters together, (ii) excluding the black shape posters from the main set of posters (i.e. using 90 posters), and (iii) using the 30 black shape posters only.

Comparing Manual and Automatic Visual Balance

The average similarity and the cosine similarity values between manual and automatic evaluation can be found in Table 1. Results for different combinations of axes and weights are presented.

By comparing the two different metrics, one can notice these return slightly different values, i.e. ranging from 0.590 to 0.807 for the average similarity and 0.900 to 0.977 for the cosine similarity metric. Even so, in one case or another, the maximum similarity values can be considered relatively high, i.e. relatively close to 1 (0.807 for average similarity and 0.977 for cosine similarity), suggesting the presented method could reasonably match the values from manual evaluation.

One can also notice that using the left-margin and bottom-margin axes to calculate the automatic balance tends to decrease similarity values when compared to the centre-vertical and centre-horizontal axes. This suggests that, in

[1]The average of the absolute differences between the manual and automatic evaluation of each poster, inverted. So the closer the value is to 1, the higher the similarity.

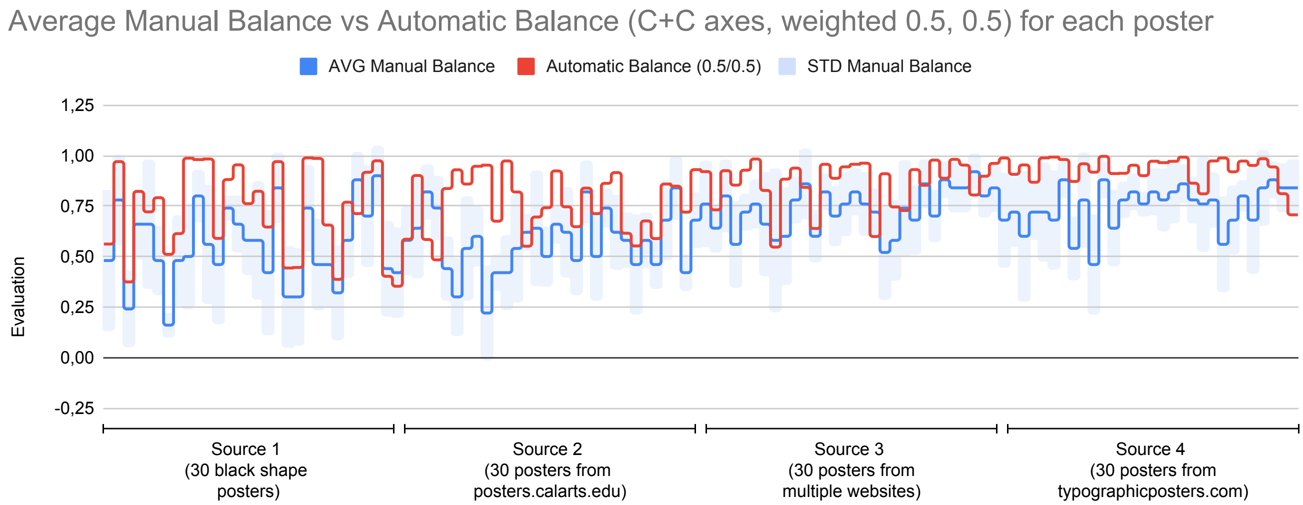


Figure 3: Automatic balance alongside Average (AVG) and Standard Deviation (STD) of manual balance, for each of the 120 posters, ordered by poster type. Automatic evaluation performed considering the C+C axes weighted equally (0.5, 0.5).

general terms, the centre-vertical and centre-horizontal axes might be better fitted to calculate visual balance. At least, when comparing to the results of the conducted survey.

Both metrics indicated the best parameterisation is using the centre-vertical and centre-horizontal axes (C+C) weighted equally (0.5, 0.5). More specifically, for such parameters, average similarity equals 0.8067 and cosine similarity equals 0.9768. Figure 3 presents a visualisation of the values obtained using the C+C axes weighted equally, alongside the average manual balance for each one of the 120 posters.

Even so, other parameterisations using one or two centre axes resulted in similar results. For instance, holding average similarities ranging from 0.788 to 0.805. This suggests that, often, the respondents evaluated better the posters in which the CM is closer to the centre of the page, e.g. either the contents align with the centre axes (one or both) or the visual weight is distributed symmetrically, relatively to the centre of the page.

Isolating the black shape posters That can also be inferred from Figure 4, which showcases all the 30 black shape posters, ordered by the respective manual evaluation values for visual balance. One can notice that most of the first 15 posters (on the top), worst evaluated, have their visual weight distributed on a single side of each centre axis (vertical and horizontal), i.e. often positioned on the corners. For instance, refer to the posters 1-8 and 10-14 of Figure 4. On the other hand, most of the posters on the bottom, better evaluated, have their visual weight better distributed on both sides of at least one centre axis. For instance, refer to posters 16 and 19-30 of Figure 4.

Best and worst-evaluated posters The assumptions above can likewise be deduced by looking at the best and worst-evaluated posters, either (a) considering all 120 posters or (b) excluding the black shape ones (see Figure 5). One can argue the visual weight of the best-evaluated posters

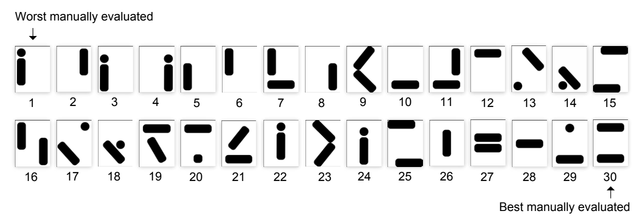


Figure 4: Black shape posters, ordered by average manual balance.

tends to be either on the centre of the page or organised in a relatively symmetrical way, relative to one or both centre axes. That can be more or less prominent in posters 2-5 of Figure 5.a and posters 2-5 of Figure 5.b. However, such an assumption can be less evident concerning poster number 1 (from either group (a) or (b) of Figure 5). Even so, one might find that the image contained in the poster reasonably balances the visual weight of the typographical elements.

On the contrary, the worst evaluated posters tend to have most of their visual weight on one side of one, or both, centre axes. For instance, all the worst posters showcased (posters 6-10 from either group (a) or (b) of Figure 5) have their contents displayed either on the left or the right of the page. Also, except for posters b.9 and b.10, all the worst-evaluated posters tend to have their contents vertically aligned to the top of the page. Poster b.9 has its contents vertically aligned to the centre of the page, and b.10 to the bottom of the page.

Differences between automatic and manual evaluation Further insights can be drawn by visualising the differences (inverted similarity values) between automatic and manual evaluation values, for each of the 120 posters. As presented in Figure 6, the average difference was 0,193 (standing for a 0,807 similarity), indicating that automatic evaluation is, on average, around 80% aligned with the opinion of the respondents of the conducted survey.

Table 1: Average similarity between manual and automatic evaluation values. Maximum value highlighted in bold.

Average Similarity

Axis	Weights (vertical-axis, horizontal-axis) n/a	0.5, 0.5	0.25, 0.75	0.75, 0.25
CV	0,7730			
CH	0,7861			
L	0,5903			
B	0,5942			
C+C		**0,8067**	0,8028	0,7966
L+B		0,5956	0,5955	0,5939
C/L+C		0,8009	0,7993	0,7889
C+C/B		0,8050	0,8013	0,7958
C/L+C/B		0,7992	0,7979	0,7880

Cosine Similarity

Axis	Weights (vertical-axis, horizontal-axis) n/a	0.5, 0.5	0.25, 0.75	0.75, 0.25
CV	0,9708			
CH	0,9620			
L	0,8999			
B	0,9179			
C+C		**0,9768**	0,9721	0,9762
L+B		0,9319	0,9310	0,9209
C/L+C		0,9753	0,9710	0,9750
C+C/B		0,9765	0,9723	0,9759
C/L+C/B		0,9751	0,9712	0,9747

Figure 5: Best and worst posters concerning visual balance according to the average manual evaluation. On the top, considering all 120 posters. On the bottom, excluding the black shape posters. The respective manual and automatic evaluation values are indicated under the respective posters.

Figure 7 showcases the posters concerning the lowest and highest absolute difference (distance) between manual and automatic balance. Although it might be difficult to draw conclusive insights from the analysis of the showcased posters, we describe some possible yet speculative reasons for the higher distances obtained (which refer to the posters at the bottom in Figure 7), i.e. why the automatic method did not match the opinion of the respondents for these cases.

The first reason relates to a known shortcoming of the presented method. In most of the reviewed posters, page items are presented in darker tones compared to the respective backgrounds. Thus, assigning heavier visual weight to darker zones usually works reasonably to asses visual balance, as previously mentioned. However, in cases in which the background is darker than the respective contents, the calculation of the CM should be (but is not so far) inverted for the CM to still be attracted in the direction of the page items, now concerning lighter tones, and not otherwise. This shortcoming can be identified in posters 6 and 7 of Figure 7, in which the system considered the CM to be almost centred despite the contents being placed on the left of the page.

Furthermore, we highlight that such an inversion should only happen, as aforementioned, if the background is darker than its contents, not whenever there are more dark pixels than light ones (or whenever the average brightness is low). For example, if a poster is almost fully filled in with black objects over white background, most pixels will be dark. However, one may perceive the black blobs (in the majority) as the items, and the white space (in minority) as the background, so the calculus shall not be inverted in this case. Hence, a more sophisticated method must be implemented in future work to distinguish between background and foreground (whenever possible) and, therefore, decide whether the calculus shall be inverted.

Although the background-detection issue might explain the high distance between manual and automatic evaluation for posters 6 and 7 (and eventually 8) of Figure 7, such an argument cannot fit, for example, posters 9 and 10.

Therefore, we believe that the apparent visual movement of the composition might also have some degree of influence on the human perception of visual balance. As an example, poster 9 is composed of 2 shapes positioned in a way the calculated CM is close to the centre of the page, leading to a high automatic balance value (0.986). However, visually, the shapes seem to be pilled in an unstable position (if making an analogy to the physical world and considering the ground to be the bottom of the poster), which might have led the respondents to evaluate this poster with a low balance value (averaging 0.460). Even so, this is a speculative assumption.

A third reason concerns poster 10 of Figure 7 and relates back to the set of axes and respective weights used to calculate balance. Although the automatic method produced identical balance values for poster 10 and its symmetrical version (i.e. 0,987), manual evaluation resulted in considerably different values. For instance, 0,460 and 0,74 for poster 10 and

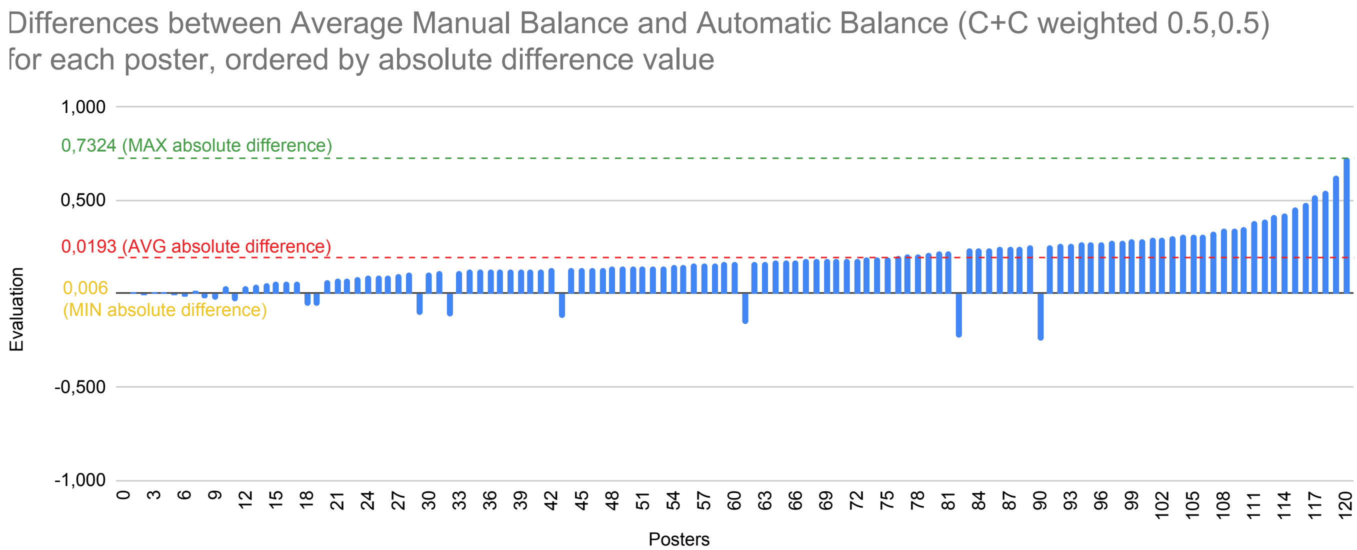


Figure 6: Differences between average manual balance and automatic balance (C+C, weighted 0.5, 0.5), for each poster, ordered by absolute difference value.

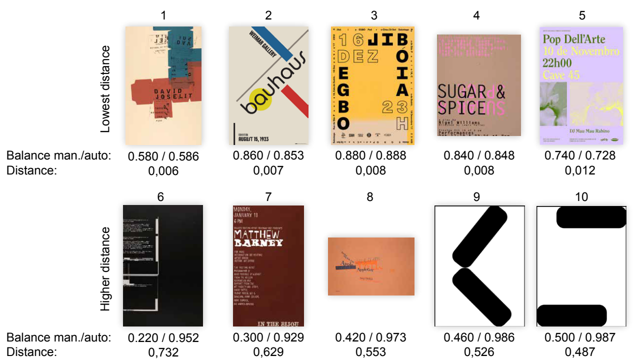


Figure 7: Posters concerning the lowest and highest absolute difference (distance) between manual and automatic balance.

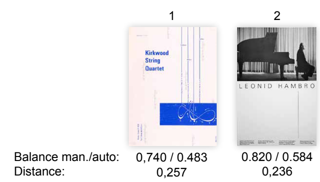


Figure 8: Two posters whose automatic evaluation is lower than manual evaluation and whose absolute difference between metrics was above average.

its symmetrical version, respectively (see posters 15 and 25 of Figure 4, respectively, for the mentioned poster and its symmetrical). A possible reason for that is the respondents considered the diagonal axis that crosses the poster from the top-left to the bottom-right corners to be more balanced than the one that goes from the top-right to the bottom-left corners. Nonetheless, all the aforementioned assumptions must require further study.

From Figure 6, one can also conclude that the automatic method often evaluates the posters optimistically compared to the average manual evaluations. For instance, 106 posters were evaluated automatically over manual evaluation, and only 14 were evaluated automatically under manual evaluation. Among the 14 under-evaluated posters, only two referred to distance values above average (see Figure 8). For instance, 0.257 and 0.236 (0.064 and 0.043 above average, respectively). Even so, it can be worth analysing such under-evaluated posters.

Although the system considered the CM was not fully centred, the respondents considered the composition of the two posters of Figure 8 to be relatively balanced, i.e. 0.740 and 0.820 manual balance, opposing to 0.483 and 0.584 automatic evaluation, respectively. A further user survey must be conducted to properly assess the reason why. Nevertheless, looking at poster 2 of Figure 8, questions concerning page division can be raised. For example, one can see that poster 2 is visually divided into two main parts — one on the top containing an image, and one on the bottom containing some typography and a wide empty zone. In future research, we shall consider whether evident divisions of the page can impact the perception of visual balance.

Conclusions In sum, considering the presented experiments and analysis, for the present experimental setup, the proposed method for automatically evaluating visual balance demonstrated to match, on around 80%, the evaluation made by the human designers and CC practitioners that participated in the conducted survey.

Therefore, we believe the present method can already be worth testing to perform fitness assignment on AEC systems. Furthermore, although the C+C axes seemed to approximate better the opinion of the respondents, we believe other pa-

Table 2: Average similarity (AVG sim.) and cosine similarity (Cosine sim.) between balance and pleasantness evaluation values, gathered through the user survey.

Poster sets	AVG sim.	Co sine sim.
All posters	0,892	0,976
Excluding black shape posters	0,923	0,991
Black shape posters only	0,800	0,934

rameterisations shall be worth trying, e.g. using C/L+C/B axes combination for allowing a wider range of possible layouts to show up.

Also, as assessing balance alone may be reductive to evaluate GD artefacts, we suggest complementing the proposed balance metric with some other metrics, such as for assessing legibility, the innovation degree of the designs or their relation to a given concept.

The conducted analysis also suggested that, presumably, other visual features can sometimes bias people's perception of visual balance. Thus, for creating more robust visual balance methods, it might be worth studying the impact of apparent movement, or how much a visual division of the page can influence balance perception.

Comparing Visual Balance and Visual Pleasantness

As mentioned before, besides visual balance, the respondents of the conducted user survey were asked, from 0 to 10, how visually pleasing they considered the posters were. The goal was to gather some insights about whether or not visual balance relates to visual pleasantness in some way. Figure 9 presents the values for balance and pleasantness gathered through the user survey, for each of the 120 posters.

Besides analysing Figure 9, to compare balance and pleasantness evaluation, the average similarity and cosine similarity were calculated. Respectively, the similarity values consisted of 0,892 and 0,976. Such relatively high values (higher than the similarity between manual and automatic evaluation) suggest there might be a considerable degree of correlation between balance and pleasantness.

For trying to gather further insights, besides the whole 120 posters, we calculated similarity values by removing the black shape posters (using 90 posters), as well as using the latter alone (30 posters only). Such values can be consulted in Table 2.

The resulting values indicate that excluding the black shape posters leads to higher similarity values (0,923 and 0,991 average and cosine similarity values, respectively). Similarly, the black shape posters alone led to lower similarity values (0,800 and 0,934 average and cosine similarity values, respectively). This might indicate the respondents found the black shape posters less visually pleasing compared to the remaining posters, regardless of their balance. Therefore, some visual features that are not as present in the black shape posters as in the remaining ones might have influenced the perception of pleasantness.

As mentioned before, we believe it might be worthy to further study what visual features influence the most the perception of visual pleasantness. Judging from the results hereby presented, one can argue that visual balance might contribute to some extent to the perception of visual pleasantness. However, further testing must be necessary to prove such an assumption.

Conclusion

One of the requirements for developing reliable and independent computational creativity systems is the ability to autonomously evaluate aesthetics. However, finding objective metrics to do it effectively is still an open problem.

In Graphic Design (GD), to properly evaluate artefacts, it may be necessary to take into account a number of factors, e.g. the visual relationship of the given artefact to its concept, how legible it is, how innovative it is, and even whether it fits the personal taste of the target audience. Furthermore, visual balance is often a relevant feature to take into consideration.

In this paper, we have presented and tested a practical method for evaluating the page balance of GD posters. To do that, a centre of mass was calculated by taking into account the brightness and location of each pixel in a given poster. The evaluation of the balance is improved by the proximity of this centre of mass to some predefined vertical and/or horizontal axes. An overall evaluation value is then determined by weighing and adding the obtained evaluation values for each axis. Different axes and combinations of axes were tested during the experiments.

To test the presented approach, a set of 120 GD posters created by different authors and gathered from various sources were evaluated manually by graphic designers and CC practitioners, by means of a user survey. The results of the survey were then compared to the ones performed by the developed method, by crossing insights from mathematical metrics and the analysis of visual features of the posters.

In addition, the respondents were asked how visually appealing they found the posters to be, hoping to retrieve some insights into a supposed correlation between page balance and visual pleasantness.

The results suggested the proposed method could match, at around 80%, the balance evaluation made by the respondents. Moreover, the results indicated a possible correlation between page balance and visual pleasantness, at least, for the current experimental setup.

Future work must focus on testing the proposed approach as a fitness assignment method for an automatic evolutionary system. As assessing balance alone may be reductive to evaluate GD posters, we must complement it with other metrics, such as for assessing legibility or innovation degree. Lastly, we must further study how different personal backgrounds or additional visual features, such as apparent movement, hue and saturation, may impact the calculation of visual balance.

Acknowledgments

This research was funded by national funds through the FCT—Foundation for Science and Technology, I.P., within the scope of the project CISUC—UID/CEC/00326/2020 and

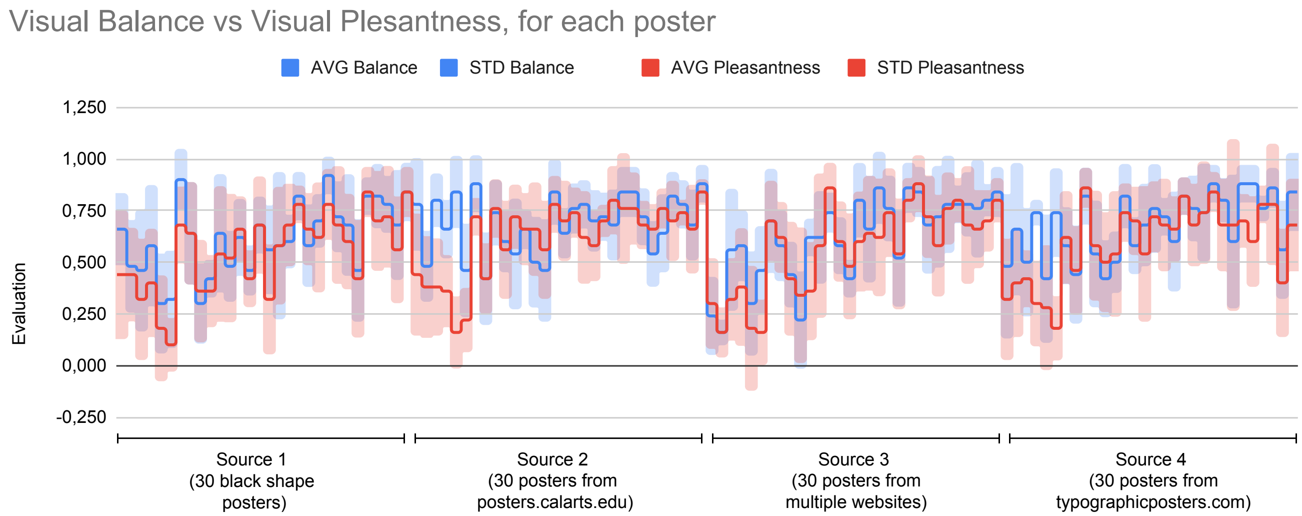


Figure 9: Average (AVG) and Standard Deviation (STD) for visual balance and visual pleasantness, for each of the 120 posters, ordered by poster type. The values were gathered through a user survey.

by the European Social Fund, through the Regional Operational Program Centro 2020 and is partially supported by *Fundação para a Ciência e Tecnologia*, under the grant SFRH/BD/143553/2019.

References

Cleveland, P. 2010. Style based automated graphic layouts. *Design Studies* 31(1):3 – 25.

Feiner, S. 1988. A grid-based approach to automating display layout. In *Proceedings Graphics Interface*, volume 88, 192–197.

Ferreira, D., et al. 2019. Design editorial algorítmico. Master's thesis, Universidade de Coimbra.

Geigel, J., and Loui, A. 2003. Using genetic algorithms for album page layouts. *IEEE MultiMedia* 10(4):16–27.

Harrington, S. J.; Naveda, J. F.; Jones, R. P.; Roetling, P.; and Thakkar, N. 2004. Aesthetic measures for automated document layout. In *Proceedings of the 2004 ACM Symposium on Document Engineering*, DocEng '04, 109–111. NY, USA: Association for Computing Machinery.

Jacobs, C.; Li, W.; Schrier, E.; Bargeron, D.; and Salesin, D. 2004. Adaptive document layout. *Communications of the ACM* 47(8):60–66.

Kitamura, S., and Kanoh, H. 2011. Developing support system for making posters with interactive evolutionary computation. In *2011 Fourth International Symposium on Computational Intelligence and Design*, volume 1, 48–51. Piscataway, USA: IEEE.

Klein, D. 2016. Evolving layout: next generation layout tool. Last accessed 17 Dec 2018.

Lok, S.; Feiner, S.; and Ngai, G. 2004. Evaluation of visual balance for automated layout. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, IUI '04, 101–108. New York, NY, USA: Association for Computing Machinery.

Lopes, D.; Correia, J.; and Machado, P. 2022. Evodesigner: Towards aiding creativity in graphic design. In Martins, T.; Rodríguez-Fernández, N.; and Rebelo, S. M., eds., *Artificial Intelligence in Music, Sound, Art and Design*, 162–178. Cham: Springer International Publishing.

Machado, P., and Cardoso, A. 2002. All the truth about NEvAr. *Applied Intelligence, Special Issue on Creative Systems* 16(2):101–119.

Martins, T.; Correia, J.; Costa, E.; and Machado, P. 2016. Evotype: from shapes to glyphs. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 261–268. ACM.

Önduygu, D. C. 2010. *Graphagos: evolutionary algorithm as a model for the creative process and as a tool to create graphic design products*. Ph.D. Dissertation, Sabanci University.

O'Donovan, P.; Agarwala, A.; and Hertzmann, A. 2014. Learning layouts for single-pagegraphic designs. *IEEE Transactions on Visualization and Computer Graphics* 20(8):1200–1213.

Rebelo, S.; Martins, P.; Bicker, J.; and Machado, P. 2017. Using computer vision techniques for moving poster design. In *6.º Conferência Internacional Ergotrip Design*. Aveiro, Portugal: UA Editora.

Rebelo, S.; Fonseca, C. M.; Bicker, J.; and Machado, P. 2018. Experiments in the development of typographical posters. In *6th Conference on Computation, Communication, Aesthetics and X*. Porto, Portugal: Uni. of Porto.

Stanley, K. O., and Miikkulainen, R. 2002. Evolving neural networks through augmenting topologies. *Evolutionary Computation* 10(2):99–127.

Zheng, X.; Qiao, X.; Cao, Y.; and Lau, R. W. H. 2019. Content-aware generative modeling of graphic design layouts. *ACM Trans. Graph.* 38(4).

Co-Creativity between Music Producers and 'Smart' versus 'Naive' Generative Systems in a Melody Composition Task

Marinus van den Oever[1,2], Anna Jordanous[3], Rob Saunders[1]
[1] Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands
[2] Zooma, Leiden, The Netherlands
[3] School of Computing, University of Kent, Canterbury, Kent, UK

Abstract

Research in computational co-creativity frequently focuses on technical performance of computational systems and subjective quality of end-products. How computational co-creative systems impact the creative process of users has received less attention. This paper reports on a two-way double-blind crossover study to investigate how the creative processes of thirteen electronic music producers were impacted while interacting with two computational co-creative systems providing melody suggestions. The two systems shared a common user interface, however, the 'smart' co-creative system suggested melodies that expanded on producers' input melodies, while the 'naive' co-creative system produced melodies unrelated to the producers' inputs. To capture participants' subjective experience, aspects of creativity were rated on Likert scales and further explored with semi-structured interviews. Each system's output and producer's intermediate melodies were compared for change (in compositions), dissimilarity (new AI-generated elements), and adoption (into melodies). Producers considered the 'smart' co-creative system to produce the most novel and valuable contributions to their process. Outputs from the 'naive' co-creative system were judged to be more dissimilar than smart expansions. Nonetheless, the changes that producers incorporated into their intermediate melodies were similar between the systems. This study suggests that co-creative interactions can stimulate the creative process by offering both related and unrelated musical suggestions.

Introduction

Computational co-creative systems have been broadly categorised into three types: creativity support tools, generative systems, and computer colleagues (Davis et al. 2015a). Computational co-creative systems have been developed to support human-computer collaboration on a range of creative tasks including drawing (Davis et al. 2015b), design (Karimi et al. 2020), games design (Yannakakis, Liapis, and Alexopoulos 2014), songwriting (Huang et al. 2020), and music improvisation (Hoffman and Weinberg 2010). But few have been developed for electronic music production (hip hop, dance, etc.). In addition, most electronic music producers use digital audio workstations (DAWs): software tools used for recording, editing, and playing back digital audio, that are yet to integrate AI technology (Davis 2022). According to Nash and Blackwell (2014), music software focuses primarily on transcribing and editing existing ideas, and not necessarily on generating inspiration. In practice, however, producers start their composition with a DAW and use it to create melodies. But current DAWs have not been designed to initiate, or help resolve creative blocks during, the composition process.

Nash and Blackwell (2014) emphasize that creativity is closely related to the transfer of ideas from the unconscious to the conscious mind, and suggest that this process can be stimulated through computational tools. In the field of design, collaborative ideation has been shown to stimulate the production of more creative ideas by exposing individuals to ideas beyond their own (Chan et al. 2017). Knotts and Collins (2020) performed a survey among music technologists, who indicated that they used tools like Magenta Studio[1] (Roberts et al. 2019) to generate ideas as a starting point for composition. Research in the use of AI in music creation has mostly focused on technical complexity (Sturm et al. 2019) and music information retrieval (Downie 2003), rather than the creative process. Some subjective user evaluations have been published (Karimi et al. 2018), and a recent study reports the (altered) experience of interactions of musicians with a keyboard player pretending to be an AI-system (Thelle and Fiebrink 2022). No studies, however, have looked at actual human-AI interactions during composition, or compared these to unaided conditions.

This paper presents a study that compared two computational co-creative music systems, a 'smart' system that processed user input and a 'naive' generator that did not. The naive generator was used as a comparator to test the assumption that any musical proposition might be helpful when a producer is in need of suggestions, regardless of how "smart" the generator is. The main hypothesis was that the smart generator will provide more valuable suggestions that are more readily incorporated into the composition; whereas the naive generator's proposals may be considered novel and surprising, but less useful.

Method

A randomized crossover study was performed in which participants were assigned to two conditions across two consec-

[1] `https://magenta.tensorflow.org/studio/`

utive sessions in double-blinded random order: co-creating with a 'smart' system and with a 'naive' system. This robust experimental setup corrects for inter-individual differences in experimental variables (e.g., comprehension of instructions, use preferences, compositional approaches, and interpretation of questionnaires and terminologies). Both systems were presented as tools to expand on a MIDI file provided by the participant. The 'smart' system considered the input when generating its expansion, whereas the 'naive' system did not. For each session of this study, participants were asked to create two melodies using their preferred DAW, while actively collaborating with one of the systems. The participants were unaware that only one of the conditions actively interacted with their submitted MIDI files. Thirteen participants were recruited through social media and personal contacts. Participants were required to have experience in producing music with software, but it was unnecessary to have a degree in music.

Generative Systems

For this experiment, pre-trained applications from Magenta Studio (Roberts et al. 2019) were modified. The smart condition interacts with Continue, which uses a recurrent neural network to expand note sequences. The naive condition works with Generate, which uses a variational autoencoder to produce melodies based on music it has been trained on. To avoid unblinding subject and researcher to the condition, both systems had identical interfaces and file sizes.

Procedure

Participating producers were requested to make two compositions with the help of the smart system on one day, and the naive system on another day, in random order. Both sessions were conducted online at the participant's home. The lead researcher and participant communicated via Zoom, which recorded the participant's voice, screen, and computer sound. The sessions were conducted either in English or Dutch. Participants were asked to think aloud during the experiment.

The session began with an explanation of the system – how the various settings work, and how to produce, include and export MIDI files. Participants received a file to install the software. They were allowed to use their preferred DAW to work on generated outputs to minimize disruption to their usual workflow. The producer was encouraged to solicit help from the generator whenever they desired by exporting and uploading their intermediate MIDI files to the generator.

When the participant was ready, the researcher shut off the video connection and no longer interfered with the experiment, but stayed online for questions. The producer then started with the assignment to create two 8-bar melodies at a tempo of 120 beats per minute (BPM) in 40 minutes while actively collaborating with the system. Additional sounds could be added to the composition if this helped the producer to get into their flow.

After the compositional assignment, participants completed a questionnaire about their demographic information, musical expertise, and experience with the system. Using 7-point Likert scales, they indicated the extent to which they found the software's output to be novel, valuable, and surprising. The interpretation of these terms was left to the participants, but their understanding was subsequently evaluated in a semi-structured interview, where participants further explained their answers. Additionally, they evaluated whether the software made idea generation easier, if it disrupted their creative process, and their intention to use it in their daily practice. These aspects were also explored further during the interview. Finally, participants were asked to submit their DAW project and generated MIDI. The first session lasted approximately 75 minutes and the second one hour.

Analysis

Video recordings and English transcripts were downloaded from the Zoom web portal. Dutch interviews were manually transcribed. For each participant and topic, a short summary of responses was made, including representative quotes, which were tabulated for further analysis and integration. DAW project files and MIDI files produced by the user and the generators were collected for the smart and naive systems. Project files were used to export the final melodies to MIDI. To allow comparison between the monophonic suggestions (single notes played at a time) made by the systems, and the sometimes polyphonic melodies (multiple notes played together) created by producers, these melodies were reduced to monophonic through manual extracting the top melody and truncation of overlapping notes.

MIDI files were analyzed using MIDI Toolbox (Eerola and Toiviainen 2004). The *meldistance* function was used to measure similarity between two MIDI files, on a scale from 0 to 1, based on the distribution of pitch classes (*pcdist1*). Using this measure three scores, *dissimilarity*, *adoption* and *change*, were obtained for each iteration i of the producer's process, where $i \in 1 \dots n-1$ and n is the number of (intermediate) melodies created by a producer. Dissimilarity, δ^i, was calculated as the average similarity between the participant's input melody, p^i, and the system's outputs based on this input, $G^i = \{g_1^i, \dots, g_m^i\}$, where m is the number of outputs generated from the same input, such that $\delta^i = \frac{1}{m}\sum_{j=0}^{m} meldistance(p^i, g_j^i)$. Adoption, α^i, is the highest similarity measure between the generator's outputs, G^i, and the producer's next intermediate melody, p^{i+1}, such that $\alpha^i = \max_{j=0}^{m} meldistance(p^{i+1}, g_j^i)$. Change, γ^i, is the similarity between p^i and p^{i+1} such that $\gamma^i = meldistance(p^i, p^{i+1})$. For further details see van den Oever (2022).

Differences in these measures between the two systems were statistically analyzed with paired two-sided Student's t-test and Fisher's exact test, with a significance level of 0.05.

Results

Thirteen male electronic music producers (mean (M) age 23; range 19-30) participated in the study. On average, they had been actively composing music for 5.5 years (standard deviation (SD) 3.3; range 2-16). Five had a formal musical education. Six considered themselves amateurs, the others

	Smart	Naive	Difference	p-value
Dissimilarity	0.459	0.737	-37.7%±9.5	**0.000002**
Adoption	0.655	0.543	20.8%±15.4	**0.0219**
Change	0.318	0.384	-17.0%±16.8	0.1857

Table 1: Average scores for dissimilarity, adoption, and change for all producers are presented as proportions of altered elements (see methods section).

were professionals or semi-professionals. The mean time spent on making music was 14.2 (SD 8.7) hours per week.

Almost all experiments went smoothly without technical difficulties. Not all subjects adhered to the instructions, however this was considered an element of artistic liberty. One producer inadvertently used the naive system in both sessions. After discovery, a third session was conducted with the smart system. Results of the two naive sessions were averaged.

Analysis of Intermediate MIDI Files

The numbers of interactions with the system ranged between 2 and 10. Although participants varied their interactions considerably between sessions (from 0 to 5), the average numbers were similar for the smart generator (M±SD 5.5±2.9) and naive system (5.3±2.2). During each interaction, producers requested between 2 and 8 melody suggestions from their generator. These requests also did not differ significantly between systems (5.5±2.2 vs 4.8±1.8, difference 14.3±2.0%, p=0.230).

Dissimilarity For all participants, the average dissimilarity scores of melodies produced by the naive system was higher than for 'smart' melodies. The difference was highly significant (p=0.000002, Table 1). This was in line with the hypothesis that the smart generator modulates on the input and will therefore return suggestions that resemble or relate to the producer's melody. In contrast, the naive system generates output autonomously, irrespective of the input.

Adoption It was expected that the 'smarter' generator would provide more useful suggestions, leading the producer to incorporate more elements of the system's suggestions in their composition. This was the case for most producers, and the difference between the two systems was statistically significant (p=0.0219, Table 1).

Change For each interaction, the melody that was fed into the system, p^i, was compared to the (intermediate) composition made by the producer, p^{i+1}. During this complex process, producers could freely incorporate musical elements generated by the system or reject the suggestions altogether. They did this to variable degrees, to follow their own flow and inspiration, or to start with an entirely new composition. The resulting changes between p^i and p^{i+1} did not differ significantly among the two generators (Table 1).

Questionnaire and Interview

Different aspects of the interaction with the smart and the naive generators were evaluated using Likert scales and in a semi-structured interview. During the interviews, many participants made comparable comments on whether they agreed or disagreed with a certain qualification of the generator. These agreements or disagreements were scored for numerical comparisons between the two conditions, using Fisher's exact test. The results of these numerical evaluations are presented in Table 2.

Value For both systems, participants generally agreed that the software outputs were valuable. The value of the smart generator was considered somewhat higher than for the naive system. The difference in Likert scores showed a trend in favor of the smart system (p=0.06766, Table 2), which evoked appreciative comments about value from most subjects. This contrasted significantly with the naive system, on which all participants gave at least one statement of disagreement (p=0.0272).

Smart Generator: Subjects largely agreed that the output of the system or the system itself was valuable. Participants frequently mentioned that the suggestions were easy to integrate. P13: "It was much better than I expected, I only had to change the timing of a single note, and it was perfect." The processing of user input allowed participants to create variations of the same melody. P8: "The idea that came out of it was quite different from what I was initially going for, but it really provided like a nice bridge from, I guess, the general vibe I was trying to create." Few disapproving comments addressed the inefficiency, as not every suggestion was equally good, requiring participants to evaluate multiple outputs. P6 describes how unfitting results can be valuable: "Even the wrong notes let you think about the possibilities."

Naive Generator: Eight subjects stated that they found the naive system of some value. Some mentioned that the gener-

	Likert Scores (M±SD)			Participants' Comments (n)				
	Smart	Naive		Smart		Naive		
			p-value	Agree	Disagree	Agree	Disagree	p-value
Value	5.62±1.19	4.77±1.48	0.06766	10	2	8	13	**0.0272**
Novelty	5.69±0.48	4.54±1.13	**0.00929**	11	2	7	8	0.0546
Surprise	5.54±1.33	5.04±1.56	0.40780	8	4	8	4	1
Idea Generation	5.31±1.44	4.58±1.26	0.16604	8	1	6	2	0.5765
Disruption	3.00±1.87	3.96±2.05	0.23732	6	5	9	6	0.4517
Daily practice	4.23±1.74	3.62±1.56	0.27461	10	2	9	7	0.2232

Table 2: Questionnaire Likert scores, and number of agreeing/disagreeing comments during interviews.

ator was most useful at the beginning of the process, to offer ideas to build upon. P5: "It did output things that I thought were useful and that I could use to make a new melody or composition." There were complaints that the system did not stick to the participant's key and rhythm. Nonetheless, after generating many results, or changing quite a bit, participants were still able to find something of value. P9: "It is productive if you're open to anything, or willing to push you in different directions, then it's definitely super valuable."

Novelty Participants largely agreed that the smart system was novel, and the naive system only slightly. The difference in Likert scores was highly significant (p=0.00929, Table 2). Comments also tended to be more supportive of novelty among the smart system compared to the naive system.

Smart Generator: Positive comments often mentioned how the system provided new insights, directions, and inspiration. Some participants appreciated the modesty of the changes suggested by the system. P6: "Although it was so simple and so minimal, it immediately gave me a new inspiration. Something I could have played myself, but didn't have in my mind at that time." Few negative comments addressed the fact that the system partially repeated their input.

Naive Generator: Participants agreeing with the novelty of the naive generator mainly talked about the dissimilarity of the output. P7: "It came with completely different things than what I imagined." Several negative comments also used the term 'randomness' to express dissatisfaction. P10: "It is a bit too random to get a melody out that works." Sometimes the naive system generated unrelated samples that were helpful. P7: "Something completely different came out, which I thought was very cool, and because of that I discarded my own piece."

Other categories Surprise, idea generation, disruption, and daily practice did not show significant results (Table 2), however occasionally helpful comments were made. The two passively collaborating systems were repeatedly stated to be perceived as 'fellow musicians'. P5 and P13 felt that the smart system provided the same effect as collaborating with human peers. Also working with the naive system approached similar stimulation. P11: "It's almost like having an additional musician who plays something in."

Discussion

We hypothesized that the smart system would be perceived as more valuable, but less novel and surprising compared to the naive generator; and that this would be reflected in higher adoption of 'smart' suggestions, and lower dissimilarity indices. The results show that participants considered the smart system more valuable and novel than the naive system. Other categories (surprise; ideation; disruption; daily practice) showed no noticeable differences. The participants' preference for the smart system is also evident in higher adoption, meaning that more elements were incorporated in the intermediate compositions. The smart output was less dissimilar compared to the naive output. There are two possibilities for this apparent discrepancy between higher value and lower dissimilarity. First, participants could have favored expansions that shared characteristics with their own input. Secondly, adoption could be higher because the smart generator repeated elements already present. Both possibilities may have contributed to the higher adoption index, but co-creative interplay also played an important role. Despite having the freedom to explore alternative melodic directions, participants tended to stick to their initial inputs even after interacting with the smart system.

The smart and naive systems both seemed to have similar and limited effects on the compositions: in both conditions, producers changed roughly 35% of their input melody (p^i) to make their next melody (p^{i+1}), see Table 1. The majority of the melodies were unchanged, suggesting that regardless of the system used, participants were disinclined to deviate too much from their ongoing composition. This could be one reason why the smart generator, which modulates on the producer's input melody, is considered significantly more valuable than the naive system (Table 2). Participants commented that the smart generator was most useful for progressing an existing composition. However, producers also mentioned that the smart generator offered value when their input (p^i) was only a few notes. Unexpected, in view of these relatively conservative preferences, the smart generator was judged to be more 'novel' than the naive system. This makes sense considering the large number of comments on how the smart system provided options that the participants did not think of.

Limitations of the study include the experimental setup, the generative systems, and the metrics used. Despite the careful design of the experiment and use of follow-up interviews, interpretations of novelty, value, and surprise between participants may still have varied, due to the inherent ambiguity of the terms (Grace et al. 2015), and thus affected the reliability of the analysis of outcomes between participants. The experimental setup was limited to comparing the naive and smart systems and did not include an unaided condition to establish how much producers alter intermediate compositions between iterations. In addition, the number of interactions with the system varied significantly between participants (2-10). The think-aloud protocol was found to be uninformative because participants were too preoccupied with the task to verbally reflect on the activity. Moreover, the generative systems used in the study were limited to the generation of monophonic melodies, and the similarity indices used do not capture all differences between melodies, e.g., rhythm and melodic contour. Furthermore, lack of diversity and gender inclusion restricts the generalizability of the results. Addressing these shortcomings opens up avenues for further work.

Conclusions

In this study, we investigated the creative processes of music producers who composed melodies, assisted by two computational co-creative systems that provided melody suggestions: a 'smart' system that processed user input, and a 'naive' generator that did not. We observed that two operationally identical systems had distinct but valuable interactions with the creative process. Participants particularly liked 'smart' system's expansions related to their own melodies, but also appreciated unexpected 'suggestions' from the naive system. Despite the systems' passive interaction, some participants felt a sense of collaboration, similar to working with other musicians. These findings highlight the importance of studying the human creative process *in situ* when developing systems for human-AI collaboration. Evaluating the system in isolation against static metrics of quality is insufficient to understand its place in a collaboration. Instead, we must consider how it integrates into the creative process. Moving forward, we encourage future studies to adopt similar methodologies that combine quantitative and qualitative metrics in a controlled blinded evaluation, to accelerate research progress in evaluating co-creative systems. To extend this research, we suggest including an unaided condition and exploring systems that can generate polyphonic melodies. Moreover, there is a need for psychometric instruments to quantify different aspects of creative processes.

Author Contributions

MvdO conducted the research, carried out the experiments, and drafted the paper. AJ and RS provided supervision and edited the document.

Acknowledgements

This paper is based on Marinus van den Oever's master thesis for the Media Technology MSc programme at Leiden University, The Netherlands. We would like to thank Oded Ben-Tal and Oliver Bown for their role as critics, providing valuable feedback during the defence.

References

Chan, J.; Siangliulue, P.; Qori McDonald, D.; Liu, R.; Moradinezhad, R.; Aman, S.; Solovey, E. T.; Gajos, K. Z.; and Dow, S. P. 2017. Semantically far inspirations considered harmful? accounting for cognitive states in collaborative ideation. In *Proc. of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 93–105. New York, NY, USA: Association for Computing Machinery.

Davis, N.; Hsiao, C.-P.; Popova, Y.; and Magerko, B. 2015a. An enactive model of creativity for computational collaboration and co-creation. In Zagalo, N., and Branco, P., eds., *Creativity in the Digital Age*. London, UK: Springer. 109–133.

Davis, N.; Hsiao, C.-P.; Singh, K. Y.; Li, L.; Moningi, S.; and Magerko, B. 2015b. Drawing apprentice: An enactive co-creative agent for artistic collaboration. In *Proc. of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, 185–186. New York, NY, USA: Association for Computing Machinery.

Davis, D. 2022. Compute and resonate: An ongoing experiment in creating acid music using accessible artificial intelligence and computer-based generative tools. In Filimowicz, M., ed., *Designing Interactions for Music and Sound*. London, UK: Focal Press. 65–82.

Downie, J. S. 2003. Music information retrieval. *Annual Review of Information Science and Technology* 37(1):295–340.

Eerola, T., and Toiviainen, P. 2004. *MIDI Toolbox: MATLAB Tools for Music Research*. Jyväskylä, Finland: University of Jyväskylä.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3–4):125–147.

Hoffman, G., and Weinberg, G. 2010. Gesture-based human-robot jazz improvisation. In *2010 IEEE International Conference on Robotics and Automation*, 582–587.

Huang, C.-Z. A.; Koops, H. V.; Newton-Rex, E.; Dinculescu, M.; and Cai, C. J. 2020. Ai song contest: Human-ai co-creation in songwriting. In *Proc. of the 21st International Society for Music Information Retrieval Conference*, 708–716.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. In *Proc. of the 9th International Conference on Computational Creativity*, 104–111. Association for Computational Creativity.

Karimi, P.; Rezwana, J.; Siddiqui, S.; Maher, M. L.; and Dehbozorgi, N. 2020. Creative sketching partner: An analysis of human-ai co-creativity. In *Proc. of the 25th International Conference on Intelligent User Interfaces*, 221–230. New York, NY, USA: Association for Computing Machinery.

Knotts, S., and Collins, N. 2020. A survey on the uptake of music ai software. In *Proc. of the International Conference on New Interfaces for Musical Expression*, 499–504.

Nash, C., and Blackwell, A. F. 2014. Flow of creative interaction with digital music notations. In Collins, K.; Kapralos, B.; and Tessler, H., eds., *The Oxford Handbook of Interactive Audio*. New York, NY, USA: Oxford University Press. 387–404.

Roberts, A.; Engel, J.; Mann, Y.; Gillick, J.; Kayacik, C.; Nørly, S.; Dinculescu, M.; Radebaugh, C.; Hawthorne, C.; and Eck, D. 2019. Magenta studio: Augmenting creativity with deep learning in ableton live. In *Proc. of the International Workshop on Musical Metacreation*.

Sturm, B. L.; Ben-Tal, O.; Monaghan, U.; Collins, N.; Herremans, D.; Chew, E.; Hadjeres, G.; Deruty, E.; and Pachet, F. 2019. Machine learning research that matters for music creation: A case study. *Journal of New Music Research* 48(1):36–55.

Thelle, N. J. W., and Fiebrink, R. 2022. How do musicians experience jamming with a co-creative "ai"? In *NeurIPS 2022 Workshop: Machine Learning for Creativity and Design*.

van den Oever, M. 2022. Co-creativity between music producers and 'smart' versus 'naive' generative systems in a melody composition task. MSc. thesis, Leiden University. Available from: https://theses.liacs.nl/pdf/2021-2022-OevervandenMarinus.pdf.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *International Conference on Foundations of Digital Games*.

Transfer Learning for Underrepresented Music Generation

Anahita Doosti and **Matthew Guzdial**
Department of Computing Science, Alberta Machine Intelligence Institute (Amii)
University of Alberta
Edmonton, AB, Canada
{doostisa, guzdial}@ualberta.ca

Abstract

This paper investigates a combinational creativity approach to transfer learning to improve the performance of deep neural network-based models for music generation on out-of-distribution (OOD) genres. We identify Iranian folk music as an example of such an OOD genre for MusicVAE, a large generative music model. We find that a combinational creativity transfer learning approach can efficiently adapt MusicVAE to an Iranian folk music dataset, indicating potential for generating underrepresented music genres in the future.

Introduction

Automated music generation has a long history (Briot, Hadjeres, and Pachet 2017). In recent years, large-scale neural network models for music generation have arisen, trained on massive datasets and requiring significant computation (Civit et al. 2022). While these approaches have proven successful at replicating genres of music like those in their training sets, due to the nature of large-scale neural network models we expect this may not prove true for dissimilar genres. Specifically, we hypothesize that these large scale models will perform poorly for out-of-distribution (OOD) genres of music, those representing underrepresented or less globally popular types of music. We therefore conducted a study on one such large-scale neural network model to understand (1) how it performed on OOD music genres, and (2) how we might best adapt the model to an OOD music genre.

For this paper, we focus on Google Magenta's MusicVAE model (Roberts et al. 2018). We lack the space for a full discussion of the model, but direct interested readers to the original paper. This hierarchical variational autoencoder is trained on an enormous dataset of roughly 1.5 million unique MIDI files collected from the web. While its exact dataset was not made public, online repositories of MIDI files are typically made up of fan-made annotations of popular songs. Thus, automatic and indiscriminate data collection would result in an unbalanced dataset in terms of genre diversity. This is due to the fact that popular chart-topping songs are much more likely to be annotated in the MIDI format. The training requirements for MusicVAE are a problem when it comes to generating underrepresented music, like experimental music or music from particular cultures with distinct musical traditions. These genres of music are unlikely to have the massive datasets needed to train models like MusicVAE. Even if such datasets existed, new musical genres are constantly being invented, meaning we could never use this approach to generate all underrepresented genres of music.

If we want to be able to generate underrepresented music, one option outside of training MusicVAE from scratch is transfer learning (Tan et al. 2018). Transfer learning refers to the collection of approaches that can adapt knowledge from a model pre-trained on some source dataset (i.e., popular MIDI files) to a target domain with limited data (i.e., underrepresented music MIDI files). However, these approaches tend to require significant similarity between the source and target domains, which may not hold true for popular and underrepresented music genres (Marchetti et al. 2021). Combinational creativity, also sometimes combinatorial creativity, is a type of creative problem solving in which two conceptual spaces are combined to represent a third or new conceptual space (Boden 2009). While different musical genres may vary in terms of their local features (e.g., melodies), they are all still music. As such, we hypothesized that a combinational creativity-inspired transfer learning approach may be able to outperform traditional transfer learning approaches at the task of adapting MusicVAE to an underrepresented genre of music (Mahajan and Guzdial 2023).

In this paper, we explore the application of CE-MCTS (Mahajan and Guzdial 2023), a combinational creativity-based transfer learning approach to adapt MusicVAE to an underrepresented music genre. While there are many deep neural network (DNN) models like MusicVAE for music generation, applying transfer learning to a DNN model for music generation remains under-explored (Svegliato and Witty 2016; Marchetti et al. 2021). In addition, while combinational-creativity-based transfer learning approaches have been applied to many domains including image classification (Banerjee 2021) and financial health prediction (Mahajan and Guzdial 2023), they have never been applied to the music generation domain. In the remainder of this paper, we first demonstrate an experiment to identify an out-of-distribution (OOD) music genre for MusicVAE. We then introduce CE-MCTS and a number of more standard transfer learning baselines. Finally, we demonstrate their performance in terms of reconstruction accuracy for an OOD

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

music dataset and present a short discussion on their music generation performance.

Related Work

There have been many recent applications of deep neural networks (DNN) to music generation (Civit et al. 2022). For instance, sequence-based approaches are popular in this field due to their ability to learn long-term dependencies in musical pieces. Multiple studies have combined sequence-based models such as Long Short-term Memory (LSTM) Recurrent Neural Networks (RNN) with autoencoders and achieved good results (Oore et al. 2017). Alternatively, Generative Adversarial Networks (GAN) have been employed to generate novel music (Yang, Chou, and Yang 2017). These models typically are also trained from scratch. However, given the data imbalance across different genres, approaches like transfer learning that can adapt knowledge from one domain to another might be useful. One such example attempts to test a pre-trained Generative Adversarial Networks (BinaryMuseGAN) with traditional Scottish music and improves its performance using finetuning (Marchetti et al. 2021). To the best of our knowledge, finetuning is the only transfer learning approach that has been applied to DNN music generation (Svegliato and Witty 2016). We use it as a baseline.

In regards to Iranian (Persian) traditional or folk music, the prior work focuses on generation of music via traditional non-machine learning approaches and/or training from scratch. In (Arshi 2018), the author uses a combination of evolutionary algorithms, Boltzmann machine models and cellular automata to generate music. They evaluate their work by the use of surveys targeted to both general and professional audiences. Alternatively, researchers have employed RNNs trained on a dataset of traditional Iranian music to generate music (Ebrahimi, Majidi, and Eshghi 2019). We note our goal is not to generate Iranian music specifically, but to explore the best ways to adapt large DNN music generation models to OOD genres like Iranian music.

Genre Analysis

MusicVAE as a music generation model boasts a very impressive performance. Specifically, it achieves 95.1% over its test dataset. However, our hypothesis was that MusicVAE would do poorly for OOD music.

To examine this question, we collected four experimental datasets of 10 songs each. These were small in size as we only required a general approximation of whether the genre was out-of-distribution for MusicVAE. We selected these songs according to two criteria: (1) if they were published after MusicVAE or were otherwise unlikely to be included in the original dataset (Roberts et al. 2018) and (2) if their genres were distinct from a melodic standpoint. Melodies can differ in many ways such as contour, range and scale and these characteristics are different across different genres (DeLone and Wittlich 1975).

Our four datasets are as follows:

- **Synth pop**, songs from a 2021 Netflix special, Inside by Bo Burnham, which musically fall into the synth pop category. This dataset serves as a comparison point, since we expected MusicVAE to perform well on this genre.

Dataset	Accuracy(%)
Synth pop	95.83
Iranian folk	43.75
Video game	84.38
Horror score	87.92

Table 1: MusicVAE accuracy on 4 datasets of different genres

- **Iranian folk**, arose from a region with a long-standing history of composing music with independent roots from western music. As such, we anticipated this would be the most challenging for MusicVAE.
- **Video game**, consists of Nintendo Entertainment System video game music. These songs have limited polyphony as only 3 notes can be played on the NES at once.
- **Horror scores**, were designed to build suspense and create a sense of foreboding. Musically this genre frequently uses dissonant notes or chords, atonality (not having a clear scale), sudden changes of tempo, and other effects to induce a sense of eeriness and dread.

In these experiments, we fed melody sequences extracted from the songs into the pre-trained MusicVAE and report the reconstruction accuracy. We include the results of our analysis in Table 1. As we expected, the model performed best on the first dataset. The 95.83% accuracy is in line with what was reported for the test accuracy on the original MusicVAE dataset. Predictably, the accuracy is noticeably lower for the other three datasets, with the Iranian dataset standing at a mere 43%, a major drop in performance compared to the rest. As such, we focused on Iranian folk music for the remainder of our study.

Iranian Folk Music Dataset

We gathered a new dataset of Iranian folk MIDI files in order to evaluate the possibility of adapting MusicVAE to this out-of-distribution (OOD) genre. This dataset consists of 100 MIDI files from both Farsi-speaking websites and from musescore.com which is a free sheet music sharing website. These files contain different instruments and varying levels of polyphony. We collected 100 MIDI files as this is in line with the target genre dataset size for prior finetuning-based transfer learning approaches with music generation models (Svegliato and Witty 2016; Marchetti et al. 2021). However, we anticipate that CE-MCTS could perform well with fewer samples (Mahajan and Guzdial 2023). During our experiments we used a five-fold cross validation, meaning we split the data into five train-test splits (80 songs for training, 20 for testing), which helped ensure that we did not just get a "lucky" train-test split.

Conceptual Expansion Monte Carlo Tree Search (CE-MCTS)

We hypothesized that a combinational creativity-based transfer learning approach could most effectively adapt MusicVAE to an OOD genre. While there have been several prior examples of combinational creativity-based transfer learning approaches (Banerjee 2021; Singamsetti, Mahajan, and Guzdial 2021), we chose Conceptual Expansion Monte Carlo Tree Search (CE-MCTS) (Mahajan and Guzdial 2023). None of these combinational creativity-based transfer learning approaches have been applied to music data, but CE-MCTS demonstrated the ability to adapt to the behaviours of distinct groups of humans across problem domains. As such, we anticipated it would be the best for adapting to other types of human expression, like music.

Here we briefly describe CE-MCTS and how we adapted it in this work. However, for a full description of the approach we direct interested readers to (Mahajan and Guzdial 2023). The "Conceptual Expansion" (CE) in the name refers to the fact that we are searching over combinations (Banerjee 2021). In this case, different combinations of the learned features from the original MusicVAE model. While it may seem unintuitive to combine features from the same model, this can allow us to approximate unseen features. This is equivalent to combining different musical patterns in the source domain (i.e., popular MIDI music) to approximate patterns in the target domain (i.e., Iranian folk music). We then employ Monte Carlo Tree Search (MCTS) to search over the space of these possible combinations. As is typical with MCTS, we build up a tree to search through this space. The root node represents the original trained MusicVAE model (no combinations) and every subsequent node represents a different set of feature combinations, with closer nodes representing more similar combinations.

For our implementation of CE-MCTS we largely employed the same setup from the original paper (Mahajan and Guzdial 2023). However, we made a number of changes for this domain. For the fitness, we used the reconstruction accuracy of the training split (80 songs from Iranian folk music dataset). We ran 10 iterations, each with 10 rollouts ($L = 5$) and $\epsilon = 0.5$. We did this to bias the search towards exploration near the original MusicVAE model as we did not want to risk catastrophic forgetting, in which a model loses useful features. Ultimately, we output the three best performing models according to the fitness and report their average performance over the test split (20 songs). The strategy for the final model selection varies based by domain.

Transfer Learning Baselines

We hypothesized that CE-MCTS would outperform standard transfer learning approaches at adapting MusicVAE to our Iranian folk music dataset. Here we introduce the transfer learning baselines we used for comparison purposes. Other transfer learning approaches were not appropriate as we lacked access to MusicVAE's original training dataset.

- **Non-transfer**, where we train a randomly initialized MusicVAE on the Iranian music dataset alone. This represents the standard approach to this problem without transfer learning. We did not expect this to work given the limited amount of training data.
- **Zero-shot**, which uses the pre-trained weights of MusicVAE with no additional training on the Iranian music dataset. We know MusicVAE does poorly when reconstructing 10 Iranian folk songs, but this won't necessarily hold for our larger 100 song dataset.
- **Finetuning (all)**, in which we use finetuning, a traditional network-based transfer learning approach that has been applied in prior music generation DNN transfer learning work (Tan et al. 2018; Svegliato and Witty 2016; Marchetti et al. 2021). In this baseline, we applied finetuning by continuing to train MusicVAE on a train split of our dataset until convergence. The (all) indicates retraining all of MusicVAE's layers. This is unusual as it can lead to catastrophic forgetting, where useful features are lost in the adaptation process.
- **Finetuning (last)**, which is the same as Finetuning (all) except that we freeze the weights of all but the last layer. This is the more typical approach when applying finetuning as it assumes that the earlier layers contain useful features (e.g., musical patterns) and we can just adapt the last layer to apply these patterns more appropriately for our target dataset.

We also developed a knowledge distillation approach called student-teacher learning (Romero et al. 2015). In this approach we trained a MusicVAE (student network) on Iranian music, through a combination of its loss and the loss of another MusicVAE with pre-trained weights (teacher network). This model proved worse than all other baselines, thus we do not include its results.

Evaluation

For our evaluation, we compared the reconstruction accuracy of our approach and baselines. Each model was fed our test set and we measured the percentage of correctly reconstructed notes. Clearly, this method of evaluation does not assess the ability of the model to actually generate new musical sequences, which is the main objective of a music generation model. However, this approach allows us to run an initial quantitative evaluation. While there are objective metrics employed by other researchers, they are not consistently defined, making it difficult to compare outputs across different generation systems. Furthermore, there is no correlation between qualitative and quantitative metrics of evaluation, making it difficult to draw implications. We expect this to be even more difficult for a genre like Iranian music or other regional genres. Therefore, correctly evaluating the quality of generation would require a human subject study with experts in the target genre (i.e. Iranian folk music). (Ji, Luo, and Yang 2020) We leave this to future work, but include a small case study below.

Results

Tables 2 and 3 contain our training and test reconstruction accuracy results, respectively. Overall we can observe that CE-MCTS consistently outperforms other methods in both training and test accuracy. Although Finetuning (last) performs similarly on average during training, CE-MCTS is better at reconstructing the test data.

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Non-transfer	68.75	68.75	68.75	68.75	68.75	68.75
Zero-shot	93.75	90.62	84.37	87.50	87.50	88.75
Finetune (all)	87.50	84.37	78.12	78.12	75.00	80.62
Finetune (last)	96.87	90.62	93.75	**100**	**100**	96.25
CE-MCTS	**98.96**	**94.80**	**98.97**	94.84	**100**	**97.52**

Table 2: Training reconstruction accuracy of each approach on each fold

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Non-transfer	37.50	37.00	37.50	6.25	53.12	34.27
Zero-shot	90.62	87.50	75.00	37.50	90.62	76.24
Finetune (all)	81.25	68.75	25.00	34.37	65.62	55.00
Finetune (last)	**96.87**	96.87	65.62	40.62	93.75	78.75
CE-MCTS	93.75	**97.9**	**83.34**	**51.07**	**93.77**	**83.97**

Table 3: Test reconstruction accuracy of each approach on each fold

As we expected, Finetuning (all) produces inferior results to Finetuning (last). In fact, it seems that the original pre-trained MusicVAE (Zero-shot) outperforms this method. This is likely due to catastrophic forgetting. As for the non-transfer method, it is not surprising that the small quantity of data available is unable to effectively train the network. The initial dataset size used to train MusicVAE is roughly 15,000 times larger than our dataset.

CE-MCTS outperforms both the pre-trained MusicVAE and last layer finetuning on test accuracy. Therefore we can deduce that by recombining the features in the earlier layers, CE-MCTS is able to create better features for the target dataset. Based on the performance of the Zero-shot and Finetuning (last) baselines, we can infer that the features present in original model are not sufficient to represent Iranian folk music. This suggests we can usefully approximate Iranian folk music features via a combination of features from popular western music.

Case Study

In this section, we provide a brief qualitative analysis of some of the models in terms of music generation instead of reconstruction. We include figures with representative outputs from the three best performing approaches. These examples were chosen by the first author, who has expertise in Iranian music, to be generally representative of the characteristics of the outputs from these approaches. This is obviously highly subjective and susceptible to confirmation bias. A study with expert participants is needed to make reliable assertions about the quality of generation, but we leave this to future work.

In each corresponding figure, the x-axis represents time in seconds, limited to 4 seconds which is the length of all 2-bar outputs by MusicVAE. The y-axis represent the pitch for the notes in the MIDI format which ranges between 0 and 127. Each red rectangle in the figure represents a continuous note.

Figure 1 represents a typical melody generated by the Zero-shot pre-trained MusicVAE. The notes in this melody sound harmonious and follow a somewhat cohesive progression. They also gradually move from a higher pitch to a lower one, spanning somewhat evenly across the melodic range (distance between the highest and the lowest pitch). Figures 2 and 3, were generated using the Finetuning (last) and CE-MCTS approaches, respectively. In the first author's subjective opinion, samples generated by these two models sound more similar to, and evocative of, the type of melodies present in Iranian folk music. This is hard to qualify but here we point out a number of characteristics commonly seen in traditional and folk Iranian (Persian) music according to (Farhat 2004) and (Iran Chamber Society 2001).

- Melodies have a narrow register (pitch range).
- Melodic movement is often achieved with conjunct steps.
- There is an emphasis on cadence, symmetry, and repetition of musical motifs at varying pitches.
- Rhythmic patterns are generally kept uncomplicated and rhythmic changes are infrequent.
- The tempo is often fast, with dense ornamentation. Similar to this, it is common to see repetitive and rapid use of the same note/pitch.

As shown in both figures, the register is more limited locally and patterns that repeat the same note appear.

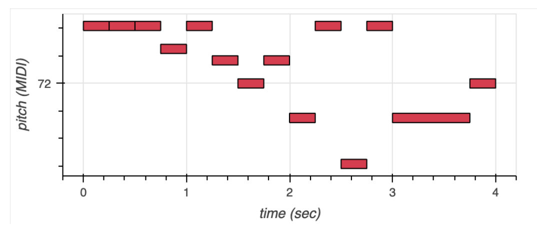


Figure 1: Visualization of a melody generated by the pre-trained MusicVAE model

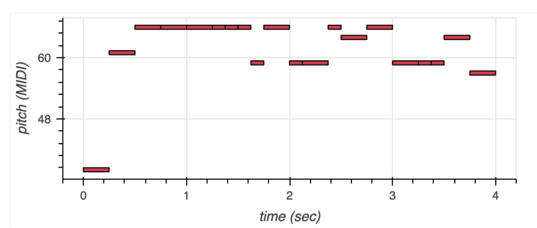


Figure 2: Visualization of a melody generated by the model finetuning the last layer

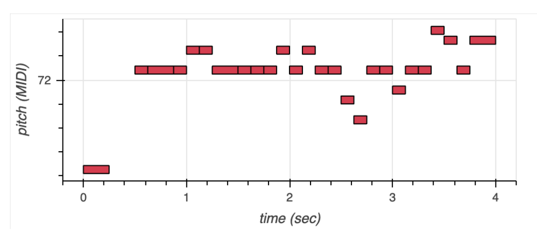


Figure 3: Visualization of a melody generated by the CE-MCTS model

Conclusions

In this paper, we investigated how MusicVAE, a music generation model, can be adapted to OOD music. We identified that MusicVAE in particular struggles with Iranian folk music. We then explored different transfer learning methods in order to improve MusicVAE's performance on a newly collected Iranian folk music dataset. Based on our results, we observed that CE-MCTS, a combinational creativity-based transfer learning approach, is better able to produce reconstructions of this genre of music. This suggests that we can successfully adapt these large music generation models for underrepresented genres of music, and that combinational creativity can be an especially helpful tool in this task.

References

Arshi, S. 2018. *Creating Persian-like music using computational intelligence*. Ph.D. Dissertation. HydraID: hull:17739 Hydra Discover Access Group: public ETD Collection: ETDComputerScience.

Banerjee, A. 2021. Combinets v2: improving conceptual expansion using sgd. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 413–413.

Boden, M. A. 2009. Computer models of creativity. *AI Magazine* 30(3):23–23.

Briot, J.-P.; Hadjeres, G.; and Pachet, F.-D. 2017. Deep learning techniques for music generation–a survey. *arXiv preprint arXiv:1709.01620*.

Civit, M.; Civit-Masot, J.; Cuadrado, F.; and Escalona, M. J. 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications* 209:118190.

DeLone, R., and Wittlich, G. E. 1975. *Melody: Linear Aspects of Twentieth-Century Music*. Prentice-Hall. 270–301.

Ebrahimi, M.; Majidi, B.; and Eshghi, M. 2019. Procedural composition of traditional persian music using deep neural networks. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, 521–525.

Farhat, H. 2004. *The dastgah concept in Persian music*. Cambridge University Press.

Iran Chamber Society. 2001. Iranian classical music. www.iranchamber.com/music/articles/iranian_classical_music.php. Accessed: April 27, 2023.

Ji, S.; Luo, J.; and Yang, X. 2020. A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions. *arXiv preprint arXiv:2011.06801*.

Mahajan, A., and Guzdial, M. 2023. *Modeling Individual Humans via a Secondary Task Transfer Learning Method*. Cham: Springer International Publishing. 259–281.

Marchetti, F.; Wilson, C.; Powell, C.; Minisci, E.; and Riccardi, A. 2021. Convolutional generative adversarial network, via transfer learning, for traditional scottish music generation. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*, 187–202. Springer.

Oore, S.; Simon, I.; Dieleman, S.; and Eck, D. 2017. Learning to create piano performances. In *NIPS 2017 Workshop on Machine Learning for Creativity and Design*.

Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; and Eck, D. 2018. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, 4364–4373. PMLR.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. Fitnets: Hints for thin deep nets.

Singamsetti, M.; Mahajan, A.; and Guzdial, M. 2021. Conceptual expansion neural architecture search (cenas). *International Conference on Computational Creativity (ICCC)*.

Svegliato, J., and Witty, S. 2016. Deep jammer: a music generation model. *Small* 6:67.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, 270–279. Springer.

Yang, L.-C.; Chou, S.-Y.; and Yang, Y.-H. 2017. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*.

Exploring Psychoacoustic Representations for Machine Learning Music Generation

Bryan Wilson, James Skripchuk, John Bacher
North Carolina State University
Raleigh, North Carolina, USA
{*bcwilso7 ,jmskripc, jtbacher*}*@ncsu.edu*

Abstract

Deep learning music generation systems have made progress in generating music artifacts ranging from scores to audio. The most successful deep learning methodologies require large amounts of computational resources, usually only available to large organizations. The environmental impact of training is non-negligible, and the computational resources can be prohibitive for research groups or independent artists engaging in co-creative design. While successful, many of these models do not take into account existing musicological domain knowledge which could yield better model performance. As a proof of concept, we augment a deep learning music generation model with an extension of a mathematical model of dissonance perception, using it to construct harmonic tension curves as an internal representation in a deep learning model. We train embeddings based on our representation and substitute them in an off-the-shelf transformer music generator. Our representation performs marginally better than baseline, with a significant reduction in training time. We explore how our representation may yield greater control of the generative space. We discuss how these results inform future research in utilizing existing domain knowledge in audio and music in order to augment deep learning models, and suggest pathways for further collaboration between computational creativity and deep learning spaces.

Introduction

Large machine learning models, particularly deep learning models that utilize many parameters to train on and generalize to a large set of data, have demonstrated incredible ability at completing complex tasks with no obvious algorithmic method. Training these large machine learning models however are computationally intensive, taking hours if not days to train. This results in high power consumption, high financial costs, and negative environmental impacts (Strubell, Ganesh, and McCallum 2020). For example, BERT (Bidirectional Encoder Representations from Transformers), a large machine learning model released by Google consisting of 110M parameters, is estimated to require 79 hours for initial training. During this initial training, BERT is estimated to consume 12kW of power, cost between $3,751-$12,570, and emit 719 lbs of CO2. (Strubell, Ganesh, and McCallum 2020). This problem affects many domains, including machine learning music generation as models such as MuseNet and Coconet consist of thousands if not millions of parameters. The size and required resources for these large models make it hard for computational creativity researchers to work with them.

Data representation has a significant impact on model training as well as the quality of music generated from machine learning models (Briot, Hadjeres, and Pachet 2017). Symbolic music representations, such as sheet music, MIDI, and chord symbols are known to decrease training time when compared to pure audio representations. Historically, work has been done on constructing additional models of how humans perceive audio and music. However, there is a lack of research on incorporating these existing theories of music perception into current representations of music, which can be useful for improving training efficiency. We aim to address this in our work.

We chose harmonic tension as our representation because it has been considered by a number of music theorists to be a strong indicator of musical coherence (Bigand, Parncutt, and Lerdahl 1996). In addition, many methods have been developed for both quantifying and modeling the harmonic tension and resolution across a piece. Specifically, we focus on the concept of a *tension curve*, a graphical model of the harmonic tension over a given chord progression (Yoo and Lee 2006). Currently, such methods are either limited to western models of music theory or only consider a finite number of chordal tones. Thus, we've designed a novel method of calculating tension curves based on psychoacoustics. To evaluate the impact of this method on the training time of machine learning music generation, we conducted a comparative study on our dataset of tension curves and a symbolic representation of music.

Data representation choice has another advantage, particularly in providing control over the generative space of the ML model. By using a representation that is suited for music similarity, for example, it is possible to take an ML music generation model, which is often seen as a black box (Castelvecchi 2016) and allow the user more control over the generative space. We perform exploratory analysis on the models output to determine the potential for greater harmonic controllability.

Related Work

Methods of Improving Training Efficiency

Current methods of improving training efficiency either fall within framework level optimization, parallel opportunities, or hardware developments (Sharir, Peleg, and Shoham 2020). Framework level optimization such as regularization and adaptive learning rate have been commonly utilized for improving model performance and training efficiency (Staib et al. 2019). However, more complex optimization approaches, such as co-designed algorithms and natural gradients have emerged more recently. Though these algorithms can lead to a quicker training time, they also can result in worse model performance (Wang et al. 2022). Current parallel opportunities, mainly within Distributed ML, are divided into two categories: data parallelism and model parallelism (Wang et al. 2022). Data parallelism requires the data to be partitioned between different nodes before fed into multiple instances of the machine learning model for training. Model parallelism requires the machine learning model be split up and placed on different devices in such a way that it can still be trained concurrently (Peteiro-Barral and Guijarro-Berdiñas 2013). While distributed ML has demonstrated success in improving training efficiency, it is very difficult to implement and more vulnerable to system failure as components are decentralized (Peteiro-Barral and Guijarro-Berdiñas 2013). Computational efficiency at the hardware level has also shown promise in improving training efficiency. There are many hardware development approaches such as memory management, dedicated hardware, and resource allocation (Markidis et al. 2018). Such approaches however have physical limitations that require constant iterations as machine learning model size increases.

Tension Curves

Even though there have been developments in expert authored music representations (Downie 2003), they haven't been utilized for machine learning music generation. One of the most notable of these representations is harmonic tension curves (Sethares 1993; Plomp and Levelt 1965; Navarro-Cáceres et al. 2020a; Yoo and Lee 2006). A harmonic tension curve models the harmonic tension and resolutions over a given piece of music by mapping a combination of tones within a chord into a single value. Common approaches are geometric mappings based on the distances between notes, such as Lerdahl's Tonal Pitch Space (Lerdahl and others 2001) or Chew's Spiral Array (Chew 2000). While useful, these do not capture any information about how humans physically perceive dissonance.

Krumhansl (Krumhansl and Shepard 1979) constructed a method where subjects were to assign a numerical rating of stability of certain pitches within a scale. While this approach takes into account human perception, it can only calculate the dissonance values of twelve notes in respect to a certain scale, and doesn't take into account the full complexity of the interaction of a note and its overtones. To mitigate this, we construct a mapping function based on an existing mathematical model of the perceived dissonance between two or more notes. To do this, we build on the approach of Vassilakis (Vassilakis and Fitz 2007), who parameterized a dissonance curve derived by Plomp and Levelt (Plomp and Levelt 1965). Not only does this allow the calculation of a dissonance value for any arbitrary collection of notes no matter the tuning or temperament, but it also includes the interaction between any arbitrary notes and their overtones.

Tension Calculation

Dissonance for Three or More Tones

In this section, we build on the work of Vassilakis to formulate a tension function able to consider a chord with an arbitrary size within the context of a piece. First, we expand Vassilakis's dissonance function to consider a chord of an arbitrary size. For chords with more than two complex tones, we calculate the dissonance of every combination of complex tones. We define D as the dissonance function developed by Vassilakis. The resulting dissonance function then becomes

$$D_v(C) = D(C_1, C_2, \ldots, C_n) = \sum_{i=1}^{Nc} D(C_a, C_b)$$

where C_a and C_b is a unique complex tone combination from the set $\{C_1, C_2, \ldots, C_n\}$ and N_c is the number of possible complex tone pairs within C.

Harmonic Tension Calculation

We will now introduce contextual components. In addition to vertical dissonance, we will also consider key tonal distance and contextual tension, as inspired by (Navarro-Cáceres et al. 2020b). However rather than use different models to calculate each component, we will be using the same dissonance function.

In regards to key tonal distance, we represent the key of our piece as a chord where each note in the key is represented in the scale. We will represent a chord representation of a key with a K where $K = [K_1, K_2, \ldots, K_n]$. Given a chord L, we will superimpose the notes of L onto the notes of K making sure to remove all duplicate notes. The dissonance therefore is calculated as

$$D_k(L, K) = D\{L_1, L_2, \ldots, L_n, K_1, K_2, \ldots, K_n\}$$

Contextual tension is based on the understanding that the perception of a chord is influenced by the chords that precede it. Similar to finding key tonal distance, we will superimpose the chord of interest onto the chord before making sure to eliminate any duplicates. Given two chords M and N with notes $[M_1, M_2, \ldots, M_n]$ and $[N_1, N_2, \ldots, N_n]$ respectively. The dissonance therefore is calculated as

$$D_P(M, N) = D\{M_1, N_2, \ldots, N_n, M_1, M_2, \ldots, M_n\}$$

Now that we have defined how to calculate every component of tension we will consider in this paper, we will now define how we aggregate these components to calculate total tension. Suppose we have a chord C_n where n is the chord position in a given piece of music in the key of K. Then we will define the total tension of chord C_n as

$$D_T(C_n) = D_v(C_n) + D_k(C_n, K) + \sum_{i=1}^{W} \gamma^i D(C_n, C_{n-i})$$

for $i \geq n$ where W is the window size and γ is the decay. Window size, W, determines how many chords before the chord of interest we consider in our contextual tension calculation. Decay, γ, determines how much our contextual tension is influenced by chords further in the past. These two values will serve as parameters to control for how much a chord's previous context influences its tension value.

Methodology

Data and Preprocessing

Our data consisted of 329 Bach chorales provided from the Music21 library (Cuthbert and Ariza 2010) designed for music analysis and processing. For every chorale in our dataset, we extracted the chords placed on the strong beats and transformed them into a list of vectors. For our tension representation, we applied our tension function to our dataset of chord vectors. For the window size parameter, W, we chose the values 1, 2, 3, 4, 5, 6, and 7 due to Bach's typical 8 beat phrasing. Since the decay parameter is confined to the range [0, 1], we chose 0.125, 0.25, 0.5, 0.75 and 0.875, in order to have an equidistant spread of values across its range. We passed our vector dataset into our tension function for all combinations of W and γ, resulting in 35 datasets of tension values. For each dataset, we allocated 80% for training, 10% for validation, and 10% for testing. Figure 1 shows a diagram of the pipeline followed for our experiment.

Training Procedure

We perform a comparative study on our symbolic representation as ground truth and our tension representation. We utilized the Music Transformer model developed by Huang et al. due to its recency in development and its manageable overhead (i.e. required training time, training data, computation power, etc) compared to other Transformer models (Huang et al. 2018). We trained our Music Transformer on the symbolic dataset using its given embedding layer and on each of our tension datasets replacing the existing embedding layer with our pretrained embedding layer. We used a batch size of 64 and trained our model for 50 epochs, each epoch consisting of 155 iterations to ensure our Music Transformer did not overfit. We used Cross Entropy Loss to evaluate the loss for each model prediction. For accuracy, we averaged the number of correct predictions across a chorale.

Results

Model Performance Analysis

We first look at the training and validation loss and accuracy curves acquired after training our Music Transformer on the symbolic dataset and our tension datasets. We only include one graph, Figure 2, for space concerns, however it depicts the benefits of our representation at a high level. Our tension representations starts at a higher training and validation loss but results in a lower training and validation loss compared to the symbolic representation. Similarly, our tension representations results in a lower training and validation accuracy but results in a higher training and validation accuracy for both compared to the symbolic representation. In addition, our tension representation converges to a lower training and validation loss and a higher training and validation accuracy quicker then the symbolic representation.

Test Dataset	Loss	Accuracy
Symbolic Control	0.671	0.823
$\gamma = 0.125$	0.667	**0.824**
$\gamma = 0.75$	**0.613**	0.818

Table 1: Best performing parameters for $W = 7$

Test Dataset	Loss	Accuracy
Symbolic Control	0.671	0.823
$W = 4$	**0.666**	0.822
$W = 6$	0.670	**0.823**

Table 2: Best performing parameters for $\gamma = 0.5$

In regards to window size, higher window sizes start with higher training and validation loss values and lower training and validation accuracy values, but end with lower training and validation loss values and higher training and validation accuracy values compared to the symbolic representation. Additionally, higher window sizes increase the rate of convergence for all loss and accuracy curves. These effects begin to diminish however for window sizes greater than 3. Decay however, had no significant effect on initial and ending values for training and validation loss and accuracy curves as well as their rates of convergence.

To evaluate the influence of our tension representation on our Music Transformer's ability to generalize to new data, we compared the testing loss and accuracy values obtained from our tension representation to that of the symbolic representation. The dataset with $\gamma = 0.75$ and $W = 7$ produced a lower testing loss compared to the symbolic representation and the dataset with $\gamma = 0.125$ and $W = 7$ produced a higher accuracy compared to that of our symbolic representation. Nevertheless, there are no significant improvements in testing loss and accuracy using our tension representation compared to our symbolic representation. Table 1 and Table 2 show testing loss and accuracy across tension function parameters $W = 7$ and $\gamma = 0.5$ respectively.

Discussion: Overall, our tension representation yields better training and validation loss and accuracy values as well as a quicker convergence time compared to the symbolic representation. This suggests that having a representation informed by human perception may allow for faster ML model training. Furthermore, using some form of intermediate representation, such as ours, to reduce training time would be beneficial for those looking to generate music with

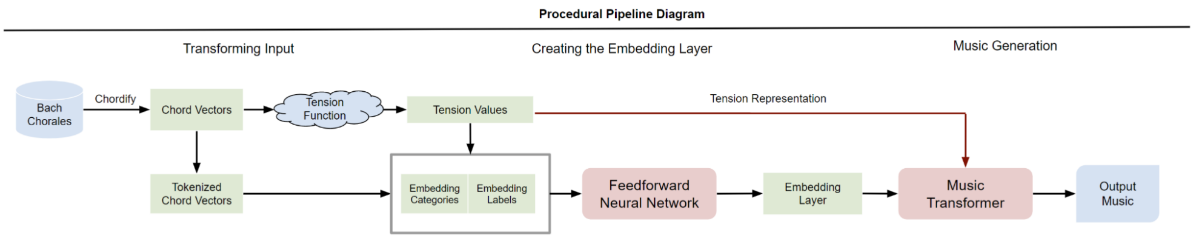


Figure 1: General Procedural Pipeline for Music Generation

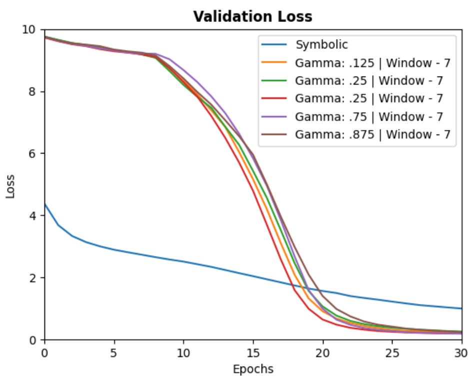


Figure 2: Validation Loss: For all γ and W = 7

limited resources.

Model Output Results

We explore our Music Transformer's output using an accepted metric of harmonic variation, Chord Histogram Entropy (CHE), proposed by (Yeh et al. 2021). To observe any correlations our decay parameter has with harmonic variation, we set $W = 7$ and calculated the CHE of both our tension representation outputs for all decay values. Our calculations exhibit a parabolic correlation, with R^2 = .875, between decay and CHE. There is not a clear correlation however, between our window size parameter and CHE.

Discussion: Our results suggest that the decay parameter, γ, has a parabolic correlation with harmonic variation. Even though we are only able to establish correlation, these results leave room for future work to determine if a causal relationship exists. Nevertheless, our model demonstrates the potential for more control and creativity focused ML models that rely on existing knowledge rather than brute-force generation. There is clearly more work to be done on making models that are sufficient for music generation tasks without the overhead of long training time and resource consumption.

Threats to Validity and Future Work

In this work, we expanded the dissonance function, proposed by Vassilakis, to incorporate both an arbitrary number of chordal tones and contextual information such as key and previous chords which we then utilized to generate a dataset of tension values to train our Music Transformer model on. Our results on model performance suggests that incorporating human perception into ML training results in higher accuracy, lower loss, and quicker training time all while producing comparable testing results. Furthermore, our results on model output explore the relationship between our tension representation parameters and the harmonic characteristics of our Music Transformer's output suggesting a correlation between contextual harmonic information and harmonic variance.

One limitation is the absence of subjective evaluation metrics such as a case study or a listening test. This makes it difficult for us to create any strong claims on the influence of our tension representation on music quality. Another limitation is that by only extracting chords on the strong beats, we limit the chord voicing range and rhythmic variance of our generated music, making it unusable for practical applications. Due to the lack of clearly detailed objective music evaluative methods, we only explored one aspects of harmonic structure, leaving many harmonic characteristics of our generated output unexplored. However due to the lack of research in utilizing psychoacoustic models for ML music generation, we believe that our limitations are valid and will be helpful for further studies in this area.

In addition to the future work that can be made to mitigate the limitations mentioned above, we used window size and decay as tension function parameters to control influence of previous chords on tension value. Our research suggests that window size influences loss and accuracy initial and ending values as well as convergence time. *What other parameters can be included in our tension function to further improve model training speed?* Furthermore, our research suggests decay exhibits a correlation to harmonic variance. However, *does this parameter influence harmonic variance?* And if so, *what other parameters can be included in a tension function to influence other music characteristics?* In addition, work has been done in performing tension curve alterations using geometric formulas to reharmonize a chord progression (Yoo and Lee 2006). *In what ways can we utilize the geometric transformation of tension curves to control harmonic interest in model output?* Finally, we only considered modeling harmony for computational music generation. *What other models can we create to influence machine learning training efficiency such as rhythm, melody, and texture through computational music generation?*

References

Bigand, E.; Parncutt, R.; and Lerdahl, F. 1996. Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & psychophysics* 58:125–141.

Briot, J.-P.; Hadjeres, G.; and Pachet, F.-D. 2017. Deep learning techniques for music generation – a survey.

Castelvecchi, D. 2016. Can we open the black box of ai? *Nature News* 538(7623):20.

Chew, E. 2000. *Towards a mathematical model of tonality*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Cuthbert, M. S., and Ariza, C. 2010. music21: A toolkit for computer-aided musicology and symbolic music data.

Downie, J. S. 2003. Music information retrieval. *Annual review of information science and technology* 37(1):295–340.

Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Simon, I.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; Dinculescu, M.; and Eck, D. 2018. Music transformer.

Krumhansl, C. L., and Shepard, R. N. 1979. Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of experimental psychology: Human Perception and Performance* 5(4):579.

Lerdahl, F., et al. 2001. *Tonal pitch space*. Oxford University Press, USA.

Markidis, S.; Chien, S. W. D.; Laure, E.; Peng, I. B.; and Vetter, J. S. 2018. NVIDIA tensor core programmability, performance & precision. *CoRR* abs/1803.04014.

Navarro-Cáceres, M.; Caetano, M.; Bernardes, G.; Sánchez-Barba, M.; and Merchán Sánchez-Jara, J. 2020a. A computational model of tonal tension profile of chord progressions in the tonal interval space. *Entropy* 22(11):1291.

Navarro-Cáceres, M.; Caetano, M.; Bernardes, G.; Sánchez-Barba, M.; and Merchán Sánchez-Jara, J. 2020b. A computational model of tonal tension profile of chord progressions in the tonal interval space. *Entropy* 22(11):1291.

Peteiro-Barral, D., and Guijarro-Berdiñas, B. 2013. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence* 2:1–11.

Plomp, R., and Levelt, W. J. M. 1965. Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America* 38(4):548–560.

Sethares, W. A. 1993. Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America* 94(3):1218–1228.

Sharir, O.; Peleg, B.; and Shoham, Y. 2020. The cost of training NLP models: A concise overview. *CoRR* abs/2004.08900.

Staib, M.; Reddi, S. J.; Kale, S.; Kumar, S.; and Sra, S. 2019. Escaping saddle points with adaptive gradient methods. *CoRR* abs/1901.09149.

Strubell, E.; Ganesh, A.; and McCallum, A. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13693–13696.

Vassilakis, P. N., and Fitz, K. 2007. Sra: A web-based research tool for spectral and roughness analysis of sound signals. In *Proceedings of the 4th Sound and Music Computing (SMC) Conference*, 319–325.

Wang, H.; Qu, Z.; Zhou, Q.; Zhang, H.; Luo, B.; Xu, W.; Guo, S.; and Li, R. 2022. A comprehensive survey on training acceleration for large machine learning models in iot. *IEEE Internet of Things Journal* 9(2):939–963.

Yeh, Y.-C.; Hsiao, W.-Y.; Fukayama, S.; Kitahara, T.; Genchel, B.; Liu, H.-M.; Dong, H.-W.; Chen, Y.; Leong, T.; and Yang, Y.-H. 2021. Automatic melody harmonization with triad chords: A comparative study. *Journal of New Music Research* 50(1):37–51.

Yoo, M.-J., and Lee, I.-K. 2006. Musical tension curves and its applications. In *ICMC*. Citeseer.

Stonkinator: An Automatic Generator of Memetic Images

José P. Lopes, João M. Cunha and Pedro Martins
University of Coimbra,
Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering
{joselopes,jmacunha,pjmm}@dei.uc.pt

Abstract

In an increasingly active and global society, memes came to be seen as a means of communication and entertainment in digital culture. Throughout the years, many image-based memes have been produced to be used in computer-mediated communication. In these images, semiotics play an important role, often involving processes of citation, parody, remix, among others. In this paper, we present a system that generates *memes* in the *Stonks* format, using a sentence introduced by the user as input. With this system, which we called *Stonkinator*, we aim to help users create their own memes. To achieve this, the system generates memes with a single input, using techniques of image analysis and blending. The *Stonkinator* system was tested by 23 participants and the results show that the system is able to produce a wide variety of memes, which, according to participants, can be used on social networks and in informal text conversations.

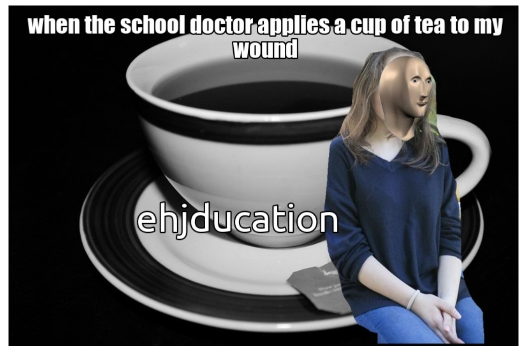


Figure 1: Meme generated by *Stonkinator* with the input "When the school doctor applies a cup of tea to my wound".

Introduction

Dawkins (1976) coined the term *meme* (from the Greek word *mimema*) to describe a transmission unit that holds an idea or behaviour between human beings, crossing generational and cultural barriers. Nowadays, the term is often associated with images and videos from the internet, usually containing humorous content, whose objective is to transmit an idea or message with the ability to cross language and cultural barriers (Shifman 2013). These image-based memes are greatly used in computer-mediated communication, being rapidly created and spread, and ultimately playing an important role in digital culture (Wiggins 2019).

The growing use of memes (Kostadinovska-Stojchevska and Shalevska 2018) has led to the development of multiple tools that help to create memes, e.g. Imgflip's MemeGenerator. Existing meme-generation systems usually allow the user to produce multiple types of output, making it difficult to determine the *genre* to which the meme belongs and understand how it should be used. In addition, different meme templates exist and not all have the same degree of complexity: some are based on a structure that has always the same background image and only the text is changed; others have a structure in which both the imagery and text are changed. In the field of Computational Creativity, there are already systems that explore meme generation related to the former kind, e.g. Oliveira, Costa, and Pinto (2016) generate memes by finding the most suitable meme template for a given news headline. The latter kind involves multimodal exploration, requiring both text and image to be changed. We consider that computational creativity techniques can be especially useful for this kind of meme generation, for example by using *conceptual extension* and *visual blending* (Cunha 2022; Cunha, Martins, and Machado 2020).

In this paper, we focus on *Stonks memes*, which are often used as reaction images, meant to portray a specific emotion in response to something that has been said, and represent an excessive proud feeling over the completion of a simple task ("Adam" and "Don" 2020). Our goal is to facilitate the creation of *Stonks memes* (Fig. 1) to be used in social media and instant message conversations. We present *Stonkinator*, a tool that takes a sentence or expression as input and uses it to generate memes in the Stonks format (Fig. 2). The *Stonkinator* system is divided into four modules: (i) the text handler; (ii) the image obtainer; (iii) the image analyser; and (iv) the image blender. The first module analyses the input sentence given by the user and extracts a theme or subject, which is sent to the second module. The second module obtains a set of images related to the theme identified by the first module. The third module analyses the images and selects one containing a person and another one to be used as the background. It also resizes the images and creates a mask image to be used in the blending process. Lastly, the fourth module receives the image containing the person, the background and the mask and performs the blending. It also writes the sentence used as input, the retrieved word related to the action, and places the Stonks character's head in the place of the person's head.

For the implementation of *Stonkinator*, we used different types of blending. In order to assess the preferred type of blending and if the users would have an interest in sharing the output obtained, we conducted a user study with 23 participants. The results show a preference for images created using a simple pasting method and positive feedback in terms of usage in private conversations and sharing on social networks. For testing purposes, *Stonkinator* was adapted to be used as a website, which we plan to make available in the future.

Overall, we consider that the presented project has the following positive aspects:

- efficiency: there's no need to manually analyse, select and segment image elements.
- diversity: by scraping multiple images related to the input, the system is able to return different results each time it runs, whether it's a different retrieved word, background or person in the output;

The remainder of this paper is organised as follows: the second section presents related work to memes and visual blending, the third section describes the *Stonkinator* framework, the fourth section describes an experiment and analyses the results obtained and the fifth section shows the improved work done to the tested version. Lastly, the sixth section presents the final conclusions and future directions.

Related Work

In this section, we introduce projects that have been a source of inspiration for the development of our work and explain how they relate to the *Stonkinator* system. We first describe systems related to meme generation then existing work on visual blending.

Memes

Due to the growing use of memes (Kostadinovska-Stojchevska and Shalevska 2018), multiple tools have been developed to help to create these images. Imgflip's MemeGenerator[1] provides a wide array of commonly used images to choose from and lets the user write their own captions with different settings, such as font family, font size, position, and more.

Another example is Imgflip's AI Memes,[2] which lets the user choose an image and generates a caption using a deep artificial neural network. A different approach is used by AI-Memes,[3] in which the user inputs a search query and is presented with 10 meme image choices that match the search criteria, from which the user selects one. Then, the AI system generates another 10 potential captions for the selected image, and the user chooses their preferred caption for the meme.

[1]https://imgflip.com/memegenerator

[2]https://imgflip.com/ai-meme

[3]https://colab.research.google.com/github/robgon-art/ai-memer/blob/main/AI_Memer.ipynb

Figure 2: Stonks meme

Different *meme templates* exist (Nissenbaum and Shifman 2018), e.g. the *Crying Peter Parker*, and these can also be divided into different *meme genres*, which are used in different situations (Shifman 2013). For example, one of the meme genres used in online forums and chats is called *reaction shots*, which are used to transmit an emotion or reaction to the addressee of the meme (Milner 2012). Existing meme-generation systems usually allow the user to produce multiple types of output, making it difficult to determine the *genre* to which the meme belongs and understand how it should be used.

Moreover, not all meme templates have the same degree of complexity: some always use the same background image and only the text is changed; others have a structure in which both the imagery and text are changed. In the field of Computational Creativity, there are already systems that explore meme generation in both formats. Oliveira, Costa, and Pinto (2016) focus on the automatic generation of internet memes, based on macro-images, that is, potential combinations between an image and a text, in order to be spread on social networks. Memes are produced from news headlines, to which, according to linguistic characteristics, certain macro-images are associated and the text is adapted according to the meme template. Another example is the work by Sadasivam et al. (2020), which also takes advantage of these templates. The authors present an automatic meme generator that creates memes based on a textual input and the system combines macro-images with text caption, using an encoder-decoder model to produce the final image. Other meme generators include the work by Lin et al. (2021), which addresses the problem of meme generation as an image captioning task by using an encoder–decoder architecture to generate Chinese meme texts that match image content. Miliani et al. (2020) propose a shared task for automatic classification of internet memes which includes meme detection, hate speech identification and event clustering. Besides these works, scholars have studied the subject of memes, including the development of other systems and models that classify (Afridi et al. 2021; Pranesh and Shekhar 2020; Singh et al. 2022; Wang et al. 2019; Yang et al. 2022) and generate (Chen et al. 2019; Peirson and Tolunay 2018; Shimomoto et al. 2019; Vyalla and Udandarao 2020; Wang and Wen 2015; Wen et al. 2015) memes. On the other hand, other meme structures involve multimodal exploration, requiring both text and image to be changed. We consider that computational creativity techniques can be especially useful for this kind of meme generation, for example by using *conceptual extension* and *visual blending* (Cunha 2022; Cunha, Martins, and Machado 2020).

Visual Blending

Considering that we focus on the generation of *Stonks memes*, which involves both image and text generation, it is important to describe existing work on visual blending.

Steinbrück (2013) developed a framework that formalises the process of conceptual blending and applies it to the visual domain. For that, the author divides the architecture of the project into five modules. The first two modules are dedicated to the acquisition of knowledge and allow a dynamic build of the base knowledge. The first module analyses the visual characteristics of the image through different computer vision algorithms and the second one gathers semantic knowledge about the concept presented in the image. The other modules deal with image composition. The third module implements the rules for the image selection, the fourth module selects the parts of each image to be involved in the image blending and, lastly, the fifth module executes the process of image blending and creates the final output.

Vismantic (Xiao and Linkola 2015) is a semi-automatic system that generates proposing images that express the meaning of an expression or sentence. It takes advantage of conceptual knowledge to find different visual representations of abstract concepts, with the ability to blend two images in three ways: juxtaposition, fusion and substitution. The system receives a sentence or expression as input and identifies the "subject" and the "message". The system itself is divided into three modules: the first one finds and filters images for the "subject" and "message", the second module analyses the images to filter them more thoroughly by predicting which set of images will create an unwanted output, the third module creates the blending between the "subject" and the "message" applying one of the three methods presented before.

Overall, our system has some similarities with the described projects but also many differences. Our system is divided into four distinct modules, similar to the framework put forth by Steinbrück, and, as such, the framework also analyses the image and isolates its constituent elements for subsequent blending. To obtain the images, an identical approach to Vismantic was used, employing the FlickrAPI library to access images from the Flickr repository, with the output being downloaded from a given input sentence. The text processing module leverages natural language processing tools to extract relevant words or expressions from the input sentence, much like the methodology described by Oliveira, Costa, and Pinto (2016).

The Stonkinator

The *Stonkinator* is a system that generates memes based on text input by the user. The resulting memes are restricted to a single format, called *Stonks*. The memes in this format are reaction images with a person and a background related

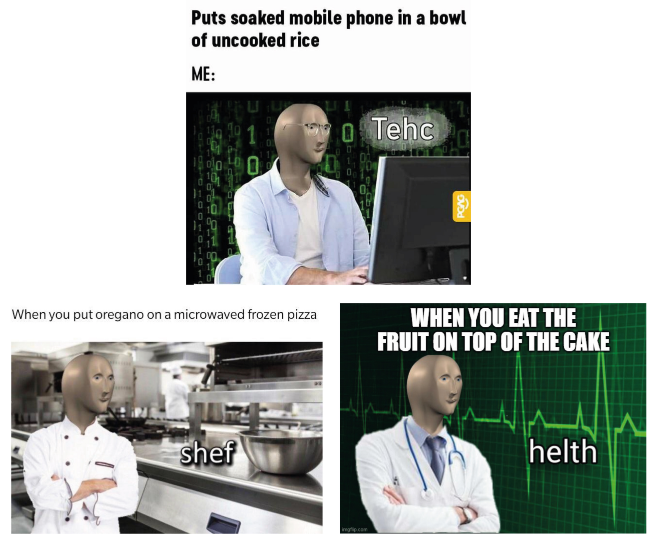


Figure 3: Example of memes in *Stonks* format

to an action described on the top of the image, in which the character is depicted as being unjustifiably proud of the action described (Fig. 2) ("Adam" and "Don" 2020).

These memes (Fig. 3) have a specific structure, being composed of four elements: (i) a character with the *Meme-man* head; (ii) a background; (iii) a sentence describing an action; and (iv) a related word or small expression. Except for the action description, which must be at the top of the image, there are no rules on the positioning of these elements, leaving this choice to the creator of the image. The *Meme-man* character head is present on every *Stonks* meme, being combined with different bodies, related to the subject of the meme. In addition, all the other elements must also have a connection with the action description subject.

This specific structure, which can be translated into a set of rules, allows us to implement a system that generates memes in this specific format. This section describes the *Stonkinator* system, which is divided into four modules (as seen in Fig. 4): (i) the text handler; (ii) the image obtainer; (iii) the image analyser; and (iv) the image blender.

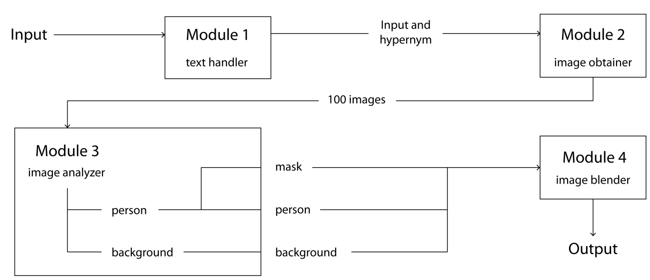


Figure 4: Stonkinator framework

In this section it is presented the system workflow through the different modules, which were implemented in Python and use several libraries for handling text and obtaining, analysing and blending images. Firstly, it will be explained how the text is handled by the first module. After that, the second and third modules are described, explaining how they work and the tools used to obtain and analyse the images. Lastly, we present the process behind the last module, responsible for the blending and composition of the final output.

Text Handling

The first module gets a sentence as input from the user and focuses on obtaining a word related to that sentence subject. The system starts by using the library RAKE (Aneesha 2015) to obtain keywords and *key expressions* with a ranking, as seen in Tables 1 and 2. It also *tokenises* the input with NLTK's (Bird, Loper, and Klein 2009) aid, a Python open-source library for Natural Language Processing, along with Wordnet (Princeton University 2010), a lexical database for the English language. Then, it selects the word or expression with the highest ranking and, in case there are multiple words, the system prioritises the selection of nouns, adjectives, adverbs, and verbs, in this order. With the word extracted, the algorithm runs different NLTK functions to search for the word's hypernyms (or hyperonyms) and synonyms. Lastly, the similarity of meanings between the chosen word and the hyperonyms and synonyms is compared, and an array is created with the words obtained ordered in descending order according to the similarity level.

Table 1: Rake keywords and ranking for the sentence "When I put a soaked mobile phone into a bowl of uncooked rice"

Keyword	Ranking
put	1.0
bowl	1.0
uncooked rice	4.0
soaked mobile phone	9.0

Table 2: Rake keywords and ranking for the sentence "When you put oregano on a microwaved frozen pizza"

Keyword	Ranking
put oregano	4.0
microwaved frozen pizza	9.0

$$\text{pizza} \begin{cases} \text{'dish'} \\ \text{'nutriment'} \\ \text{'food'} \end{cases}$$

Figure 5: Words obtained from 'pizza'

Image Retrieval and Analysis

Using the first word from the array created before as a search query, the second module downloads one hundred images from the Flickr database into a temporary folder. Then the first part of the third module analyses those saved images and identifies faces and human figures, using OpenCV and Dlib. First, it searches for faces using Dlib's HoG Detector and separates the images with or without a face detected into two different folders ("backgrounds" and "faces"). After that it searches for human figures, in the *faces* folder, using a pre-trained model trained on the COCO dataset and a library called ImageAI,[4] returning a rectangle's position that contains the human figure and a percentage of the prediction (Fig. 6a), meaning that the highest the percentage, the highest the probability of the detection is an actual person. To be easier to analyse and segment the image elements, the system uses the returned rectangles, previously mentioned, and extracts each one into a new image. Lastly, the system uses MediaPipe's tools *face mesh* and *pose* to detect the orientation of the person's face and whether it was a full-body picture, a close-up or a medium-range shot. This allows the system to execute an early filter and remove images not suitable to use, such as images without human figures or with a head orientation not suitable to place the *Mememan*'s head.

(a) Human figure detection using ImageAI

(b) Mask created using OpenCV and Mask R-CNN

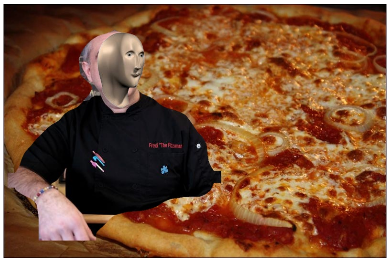
(c) Overlay of the mask and image

Figure 6: Outputs of different libraries for the image selection and segmentation

Although this process uses many techniques to select the final images, it also boosts its efficiency and reduces the error margin, by doing a wide filter first, using a fast algorithm but more probable to return false positives, and a more thorough one later, with a more complex model that can give accurate results on a smaller set of images. Also, by deleting the images that don't satisfy the requirements for the final blend during this process, it allows the system to be faster when creating another image with the same input, since the images have already been filtered and checked, and that process won't be repeated to every image when creating a new output.

After separating and filtering the images, if the *background* or *faces* folders are empty, the system downloads another set of one hundred images for the next word in the array returned by the first module and runs the first part of the third module again. This loop continues until both fold-

[4] https://github.com/OlafenwaMoses/ImageAI



(a) Paste method

(b) Laplacian method

(c) Poisson seamless cloning method

Figure 7: Outputs of different blending methods

ers contain usable images. For the image selection, in the first run of a determined input, the system selects a random background and the image with a higher probability of containing a human figure, and on the following runs of the same input, the system selects a random image from both folders to return a different result.

Image Blending

With the chosen images, the third module creates a *mask* (Fig. 6b and 6c) for the human figure, using OpenCV and Mask R-CNN architecture (He et al. 2017), and resizes the *mask* and *human* images to match the background size, using Python Image Library and OpenCV. Then it sends the *background*, *person* and *mask* images, along with the input and the first word of the array from the first module, to the fourth and last module. Firstly, this module pastes the person segment into the *background* image, using the *mask* from the previous module. For the purpose of testing, at this stage of the process, there are three different blending algorithms implemented. The first and the simplest is just pasting the human figure in the background (Fig. 7a). The second one is the Laplacian Blending (or Pyramid Blending) using five levels (Fig. 7b). The third algorithm used is the Poisson Blending,[5] using the seamless cloning method (Fig. 7c). After that, the system uses the same techniques to detect the position of the face and human figure and paste the *Mememan*'s head in the position of the human head, being the direction of the head determined with MediaPipe's *face mesh* function. Detecting both the face and human positions, instead of getting only the face location, before pasting the *Mememan*'s head, helps prevent the blending on inexisting faces detected by the algorithm.

Lastly, the final image composition is created by adding the action description at the top of the image and pasting the returned word on the final output (as seen in Fig. 7) with the help of an MSER detector (Matas et al. 2004), which returns regions of the image to place the word. During the testing phase, the font type was the default and the font size was fixed for both the description and the returned word (Figures 8, 9, 10, 11). For the current state of the system, the output comes with its caption and text written in different fonts and sizes, being this process specified in the Improved Work section.

[5] https://github.com/rinsa318/poisson-image-editing

Figure 8: Meme generated with the input "When you get early so you don't miss the bus".

Figure 9: Meme generated with the input "When you get early so you don't miss the bus".

Web interface

In order to test *Stonkinator*, we developed a web interface using Python's Flask. The Flask app has an input bar on the homepage that allows the user to write an expression or sentence. Then, that input is sent to the Python script and the process described above is started. While waiting, the user can check all the saved memes created by the system in a slideshow format.

At the end of the process, the system gives feedback noise, indicating that the output was delivered, and the user is presented with three memes, one for each blending method, using the same images selected by the third module. On this webpage, there are also three buttons. The first one simply returns the user to the homepage. The second button allows the user to generate another image with the same input. This process is faster than generating a new meme with a different input since the images were already filtered and can all be used in the new image generation process. The last button saves the images obtained into the system, so they can be seen by anyone in the gallery or the homepage slideshow. Although the *Stonkinator* webpage was only available for access during the testing period, our goal is to make it available in the future.

Figure 10: Meme generated by a participant with the input "I like to eat marmalade with cheese".

Figure 11: Meme generated by a participant with the input "Me after checking 2 + 2 on the calculator".

Experimentation

This section presents and discusses the experimental results. We begin by describing a user study and its results. Then, we provide a general analysis of the system and its output.

We conducted a user study to assess the quality of the system in terms of its technical results,[6] preferences for blending methods, predictability, usability, and utility. The main goal was to understand if the output of the system was considered a *meme* from the perspective of the users, if the resulting meme was a predictable output in the sense that the output transmits the message the user wants to convey, how easy it was to use the presented system, if the users would use the output in social media and instant messages, which blending method is preferred by the users and what is their feedback regarding the meme generated.

Experimental Setup

To obtain this data, the participants were presented with a couple of tasks and were asked to answer a series of questions related to the tasks they performed and the results obtained. The testing process started by providing users with access to the website and asking them to create a meme by introducing a sentence as input, created by the participants (e.g.: "I like to eat marmalade with cheese", Fig. 10). After that, they were presented with a set of three images, each image blended with a different method, and next, the second task was to save the set into the system if they found the meme interesting, to be displayed in the website's gallery. With this step completed, the third task asked the participants to generate another meme using the same input, and once again, the next task was to save the generated set if they wanted to. In the end, they were given the option of answering a questionnaire or continuing to use the system freely and answering later. Before answering any questions, the participants were asked to choose a set of images that they created and open it in a new window to use later in the questionnaire. The questionnaire was designed using Google Forms with the goal of studying the topics already mentioned. Users were guided through the tasks and any doubts and questions were clarified to avoid misunderstandings. The questionnaire was composed of 11 tasks:

- T1: Evaluate the technical results of the output between 1 (bad quality) and 10 (good quality)
- T2: Evaluate the predictability of the output between 1 (unpredictable) and 10 (predictable)
- T3: Evaluate the usability of the system between 1 (confusing) and 10 (intuitive)
- T4: Describe the generated meme (this was an open-ended question, even though some examples were given to the user, e.g.: funny, boring, non-sense, ...)
- T5: Would you share the meme among friends/on social networks
- T6: Would you share the meme during an informal text conversation

For the last four questions, the participants had to analyse different sets of images, three created purposely for the questionnaire and one created by them, and identify the preferred blending type. For this, we chose three generated memes (each being a set of images produced using the three types of blending) and asked the participant to conduct the following task for each meme (T7-9): Choose the preferred blending method among the set of images presented.

Then, the same task was asked for the chosen set of images selected by the participant during the system testing (T10). The last task (T11) concerned an optional open-ended question asking for comments and feedback about the test and the system.

Results

In total, *Stonkinator* was tested by 23 participants (aged between 17 and 25). To better show the results, in Table 3a, we divided the answers into three groups: "bad" (answer of less than 6 out of 10); "ok" (answer of 6 or 7 out of 10); "good" (answer of more than 7 out of 10).

Overall, the analysis of the experimental results indicates that although the average value of the evaluations regarding the technical quality of the images obtained is generally good, the same cannot be said about how predictable the generated meme is, according to Table 3a. However, the

[6] Image analysis, image segmentation, image blending and word obtained

Table 3: Results of testing

(a) Results of evaluating different parameters

Evaluation	bad	ok	good
Technical Quality	4.35%	30.43%	65.22%
Predictability	68.57%	21.74%	8.69%
Usability	0%	34.78%	65.22%

(b) Uses cases for the output

Evaluation	yes	no
Share in social media	91.3%	8.7%
Use in private conversation	100%	0%

(c) Preference on the blending method

Method	Percentage
Paste	69.56%
Laplacian	13.04%
Poisson	17.4%

participants related that these low values in terms of predictability did not impact the transmitted message. As for usability, the result is generally good, with only two people dissatisfied with the feedback obtained when saving images. Only two users reported that they would not share the result obtained on social networks, but one hundred percent of respondents would use the meme obtained in informal text conversations. Finally, regarding the blending methods, according to Table 3c, we can see that the method with the most appealing results for users is simple pasting.

General Discussion

One of the biggest criticisms regarding the technical quality of the generated memes was the image segmentation part, where it failed to detect accurately the borders of human figures in the image, being, therefore, a starting point to improve the system. A fact that may be in favour of these flaws is related to the existence of memes whose objective is to have these same flaws intentionally displayed on the image,[7] although it's not the purpose of the system. This fact may also explain why users tend to prefer simpler pasting blending methods, rather than more complex ones, according to Table 3c, since several memes were made by simply pasting elements on an image. In Table 3a we can check that the predictability of the image obtained is very low. Since the purpose of the system is to create an image capable of conveying an idea or showing a reaction close to the one intended by the user, but also create different images each time the system produces an output, we can observe that the system often produces images with a high level of unpredictability. Users reported that the generated output conveyed the message they wanted to transmit. Taking this into consideration, we interpret that the high level of unpredictability may be related to the wide variety of results for the same input, which may indicate some degree of creativity. We also consider that the high level of unpredictability is not an issue since, as seen in Table 3b, the participants stated that they would use the produced memes in a social media post or an informal text conversation. Lastly, according to user feedback, another important element to be improved was the typography. Both sentences were sometimes hard to read on a smartphone and sometimes even on a computer, as such, it's a detail that needed to be addressed.

[7] https://knowyourmeme.com/memes/dank-memes/

Figure 12: Meme generated with the input "When you put oregano on a microwaved frozen pizza".

Figure 13: Meme generated with the input "When you put oregano on a microwaved frozen pizza".

Despite the issues pointed out by users, we found that most of them managed to use the system without much difficulty, achieving the goal of being a simple "one input" system, and that more than ninety percent of the users would share the meme produced on a social network, being that one hundred percent would use it in an informal text conversation. This shows that, even if the final output has minor issues, the system can be used to achieve the results for which it was proposed. Throughout this paper we show memes produced by our system. We can see two different images generated with the same input, in Figures 8 and 9, using the default font, and in Figures 12 and 13. Other images created with different inputs are shown in Figures 10 and 11. Some of these memes (Fig. 10 and 11) were produced by our participants during the testing period.

Improved Work

After testing, by taking advantage of the participants' feedback, the *Stonkinator* was improved, more specifically its typography and the way it handles the written word. In addition, by analysing both other Stonks memes and memes

Figure 14: Meme generated with the input “When the school doctor applies a cup of tea to my wound”.

Figure 15: Meme generated with the input “When you restart the router at your parents house”.

in different formats, it was observed that some key features could be added to make the output more appealing. The improved work was made iteratively and its evolution can be seen in Figures 9, 12 and 16, corresponding to the first, second and third iteration of *Stonkinator*, respectively.

Typography

One of the problems related to the output was related to the way it wrote text in the image. Using a fixed font size for the caption and the retrieved word meant that sometimes the text could be hard to read, being perceived as a lot smaller than it was intended. This coupled with the default font having a low weight (e.g. Fig. 10) made us search for another way to display the meme caption and the selected word. Firstly we tackled the retrieved word, by changing its font to one easier to read and changing its size. The chosen font was Ubuntu-Regular for being similar to the example memes (Fig. 3) while being slightly different, to serve as a signature of the *Stonkinator*. The word size is changed to be proportional to the area returned by the MSER detector. In the first iteration, we also changed the caption font to Ubuntu-Regular but we thought that it could still be improved. By searching and analysing other types of memes, we found that a good alternative would be to use the Impact font that is present in most of the memes (Milner 2012). With this change in mind, we also found that the system would benefit from a variable font size for the caption, which was implemented to be proportionate to the size of the image.

Figure 16: Meme generated with the input “When you restart the router at your parents house”.

Stonks Deliberate Mistake

Lastly, a key characteristic of the Stonks meme is a deliberate mistake in the written word. For the original meme, the word *stocks* is changed to *stonks*, and in the examples is: *chef* to *shef*, *health* to *helth*, and *tech* to *tehc*. We also wanted to include this deliberate error and came up with a solution using phonemes. To achieve this objective, we used a Python library called Pronouncingpy[8] that returned an array containing the word phonemes. Then, unnecessary information from the array was removed and it was converted into a string that represented the word in phonemes only. For the second part, the selected word for the meme and the phoneme string were divided into syllables, using FinnSyll.[9] From here, the returned arrays were compared and a syllable from the word is changed to its respective phoneme, creating a slight mistake in the written word that keeps enough information to be understood (Figures 1,14, 15 and 16).

The results show that this approach can produce results similar to the examples above, by changing a letter or by adding or subtracting letters. Some of these changes can produce better results, like *worker* to *werker*, while others may not work as well, like *learning* to *lerarning*. This approach also introduced new ways of incorporating typos without undermining the meaning of the word by adding or changing letters to maintain how the word sounds, for example, *device* to *dihvice* and *education* to *ehjducation*. Nonetheless, although the results seem promising, further testing with users is still needed.

Conclusion

In this paper, we present a new system, *Stonkinator*, capable of creating memes in a specific format with a single input. A system was developed using multiple Python libraries, including methods and algorithms for text analysis, image analysis, image segmentation, and image blending. The system was deployed into a website that presented the output using three different blending methods. We conducted a user study with the goal of understanding the audience’s blending methods preferences and the viability of the system. The results show that our system is able to create satisfactory images, being the most visually appealing the ones with a simple pasting blending method, and that users would share the

[8]https://github.com/aparrish/pronouncingpy
[9]https://github.com/tsnaomi/finnsyll

generated memes on their social networks and in the context of a private text conversation. The system was also later improved, changing its typography and adding a small typo, characteristic of the Stonks memes. Having room for improvement, in the future we would like to improve the described approach by implementing a better image segmentation algorithm, as well as creating a better image selection system, so the output of the machine can better meet the expected ideas the user wants to convey.

Acknowledgements

This work is funded by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020

References

"Adam", and "Don". 2020. Meme man wurds / stonks edits. https://knowyourmeme.com/memes/meme-man-wurds-stonks-edits.

Afridi, T. H.; Alam, A.; Khan, M. N.; Khan, J.; and Lee, Y. K. 2021. A multimodal memes classification: A survey and open research issues. *Lecture Notes in Networks and Systems* 183:1451–1466.

Aneesha. 2015. Rake. https://github.com/aneesha/rake.

Bird, S.; Loper, E.; and Klein, E. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Chen, Y.; Wang, Z.; Wu, B.; Li, M.; Zhang, H.; Ma, L.; Liu, F.; Feng, Q.; and Wang, B. 2019. Memefacegenerator: Adversarial synthesis of chinese meme-face from natural sentences. *arXiv:1908.05138*.

Cunha, J. M.; Martins, P.; and Machado, P. 2020. Let's figure this out: A roadmap for visual conceptual blending. In *Proceedings of the Eleventh International Conference on Computational Creativity*.

Cunha, J. M. 2022. *Generation of Concept Representative Symbols: Towards Visual Conceptual Blending*. Ph.D. Dissertation, University of Coimbra.

Dawkins, R. 1976. *The Selfish Gene*. Oxford University Press.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

Kostadinovska-Stojchevska, B., and Shalevska, E. 2018. Internet memes and their socio-linguistic features. *European Journal of Literature, Language and Linguistics Studies* 2(4).

Lin, W.; Qimeng, Z.; Kim, Y.; Wu, R.; Jin, H.; Deng, H.; Luo, P.; and Kim, C. H. 2021. Automatic chinese meme generation using deep neural networks. *IEEE Access* 9:152657–152667.

Matas, J.; Chum, O.; Urban, M.; and Pajdla, T. 2004. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing* 22(10):761–767.

Miliani, M.; Giorgi, G.; Rama, I.; Anselmi, G.; and Lebani, G. E. 2020. Dankmemes @ evalita 2020: The memeing of life: Memes, multimodality and politics. In *CEUR Workshop Proceedings*, volume 2765. CEUR-WS.

Milner, R. M. 2012. *The world made meme: Discourse and identity in participatory media*. Ph.D. Dissertation, University of Kansas.

Nissenbaum, A., and Shifman, L. 2018. Meme templates as expressive repertoires in a globalizing world: A cross-linguistic study. *Journal of Computer-Mediated Communication* 23(5):294–310.

Oliveira, H. G.; Costa, D.; and Pinto, A. M. 2016. One does not simply produce funny memes!–explorations on the automatic generation of internet humor. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016). Paris, France*.

Peirson, A. L., and Tolunay, E. M. 2018. Dank learning: Generating memes using deep neural networks. *arXiv:1806.04510*.

Pranesh, R. R., and Shekhar, A. 2020. Memesem:a multi-modal framework for sentimental analysis of meme via transfer learning. In *4th Lifelong Machine Learning Workshop at ICML 2020*.

Princeton University. 2010. "About WordNet". WordNet. https://wordnet.princeton.edu/.

Sadasivam, A.; Gunasekar, K.; Davulcu, H.; and Yang, Y. 2020. memebot: Towards automatic image meme generation. *arXiv:2004.14571*.

Shifman, L. 2013. *Memes in Digital Culture*. The MIT Press.

Shimomoto, E. K.; Souza, L. S.; Gatto, B. B.; and Fukui, K. 2019. News2meme: An automatic content generator from news based on word subspaces from text and image. In *2019 16th International Conference on Machine Vision Applications (MVA)*, 1–6. IEEE.

Singh, B.; Upadhyay, N.; Verma, S.; and Bhandari, S. 2022. Classification of hateful memes using multimodal models. *Data Intelligence and Cognitive Informatics* 181–192.

Steinbrück, A. 2013. Conceptual blending for the visual domain. *Ph. D. dissertation, Masters thesis*.

Vyalla, S. R., and Udandarao, V. 2020. Memeify: A large-scale meme generation system. In *ACM International Conference Proceeding Series*, 307–311. Association for Computing Machinery.

Wang, W. Y., and Wen, M. 2015. I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 355–365. Association for Computational Linguistics.

Wang, L. Z.; Zhao, Z. D.; Jiang, J.; Guo, B. H.; Wang, X.; Huang, Z. G.; and Lai, Y. C. 2019. A model for meme popularity growth in social networking systems based on biological principle and human interest dynamics. *Chaos* 29.

Wen, M.; Baym, N.; Tamuz, O.; Teevan, J.; Dumais, S.; and Kalai, A. 2015. Omg ur funny! computer-aided humor with an application to chat. In *Proceedings of the Sixth International Conference on Computational Creativity*.

Wiggins, B. E. 2019. *The discursive power of memes in digital culture: Ideology, semiotics, and intertextuality*. Routledge.

Xiao, P., and Linkola, S. M. 2015. Vismantic: Meaning-making with images. In *Proceedings of the Sixth International Conference on Computational Creativity*. Brigham Young University.

Yang, F.; Qiao, Y.; Qi, Y.; Bo, J.; and Wang, X. 2022. Bmp: A blockchain assisted meme prediction method through exploring contextual factors from social networks. *Information Sciences* 603:262–288.

Neuro-Symbolic Composition of Music with Talking Points

Simon Colton,[1,2] **Berker Banar**[1] and **Sara Cardinale**[1]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]SensiLab, Faculty of Information Technology, Monash University, Australia
s.colton@qmul.ac.uk

Abstract

We describe the Parley system, which generates musical compositions with the process foregrounded via talking points that help users engage. Parley employs a neuro-symbolic approach combining rule-based modules and pre-trained neural models to produce music in a communicable way, akin to standard composition approaches involving designing, writing, listening to, editing and analysing music. We highlight the potential of the system in a case study where users employ Parley's modules in a Colab notebook. To investigate the interplay of the rule-based and neural processes, we describe some experimental results comparing generated music before and after an editing process which employs a neural listening model.

Introduction and Motivation

There are many reasons why people compose music, only one of which is to have more music of high quality. We take inspiration from the YouTube broadcasts of well-known classical music composer David Bruce (youtube.com/@DBruce), where composition techniques are explained for educational and entertainment purposes, via the practice of composing new pieces of music, or explaining the processes leading to existing ones. In one episode, for example, he helps a novice composer to improve their first composition, offering numerous pieces of advice while composing an enhanced version, for the educational/entertainment benefit of the novice and the viewers on YouTube. In general, such advice on composition often falls into one of three categories:

• Following rules and heuristics of music theory and practice. Advice of this nature appeals to established rules based on agreed upon concepts, often phrased in terms of constraints, requirements and best practice.

• Making choices based on listening to the emerging composition. Choices don't have to be rule-driven, but rather composers are advised to try some alternatives and choose which sounds the best to them, in terms of the rest of the composition, or a particular genre or style, or general musicality.

• Striving for global properties of the composition. Strictly following rules can lead to tedious and repetitive music, so suggestions for introducing novelty and variety are often given. In contrast, composed music can sometimes lack coherence and structure, so advice on how to achieve consistency and regularity is also given.

We are building the *Parley* generative music system to compose music with a relatively transparent process which enables users to understand its process in a way that could potentially be entertaining or educational or both. We introduce the notion of a *talking point* as some piece of information that Parley can communicate in terms of rules employed, aesthetic preferences determined through listening or striving for a global property of the music it is generating. The purpose of a talking point is to communicate a decision made during a composition process or an opinion about an excerpt of music which provokes the user to engage with the music/process by possibly agreeing or disagreeing with Parley's opinion.

We have argued in (Colton and Banar 2023) that, while generative deep learning is the dominant force currently and undoubtedly produces highly impressive music (for instance see (Agostinelli et al. 2023)), it is not necessarily the best option if there are other reasons – such as education and entertainment – for the composing of music. This is, in part, because the black-box nature of deep learning makes it difficult to foreground decisions which have been made in the generative process. We suggest instead a neuro-symbolic (Hitzler et al. 2022) approach which employs rule-based and logical AI techniques in tandem with neural models to have both generative power and communicable decision making. With particular application to composition of music with talking points in mind, Parley employs rule-based approaches to **generate** initial compositions and neural *listening models* to estimate listener expectations and to tag excerpts in order to **edit** an evolving composition, as well as to **design** the overall form of the composition and **analyse** the music produced.

In future work, we will compare Parley with relevant prior work in computer assisted composition, including the PatchWork and OpenMusic projects (Assayag et al. 1999). We concentrate here on describing how Parley operates. In the next two sections, we describe the listening models employed, and the neuro-symbolic process in terms of a modular approach where users can employ multiple designer, generator, editor and analyser modules. We follow this with a case study, where a user co-creates compositions with Parley in the 'Flaneur' series, via multiple stages in a Colab notebook employing Parley's modules. We then describe some experiments which investigate the nature of the music produced with and without editing enabled by a listening model. We end by describing the general potential for neuro-symbolic approaches in computational creativity projects, and highlighting some future directions for the Parley project.

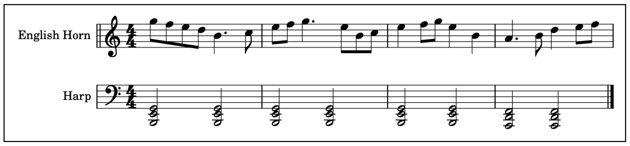


Figure 1: Example excerpt for the mood tagging experiment.

Neural Editing and Listening Models

Parley uses pre-trained neural models from two different projects, to perform functions comparable with human composers listening to their composition, as described below.

A Melody Pitch Class Prediction Model

It is standard to ascribe *the pitch class* of the note middle C – and any note which differs by this by a multiple of 12 semitones (an octave) – to be 0, with C# ascribed 1, D ascribed 2, and so on up the semitone scale. As described in (Banar and Colton 2022), an LSTM neural model (Hochreiter and Schmidhuber 1997) was trained to be able to predict the *interestingness* of each note's pitch class in a given piece of music. That is, the LSTM ascribes a score for each of 11 pitch classes, based on a window of notes preceding it. The more *unlikely* an observed pitch class is, according to the activations of the model in the output layer, the more interesting it is, and this process was used to help identify decision points in composed music. Two models were trained, on melodies written in the 18th and 20th Centuries respectively.

We repurposed the former of these models for Parley to use to edit the pitches in a melody to better fit the training set distribution of this pre-trained model better (i.e., with higher likelihood, lower interestingness). This could help produce melodies that adhere to traditional classical music expectations rather than more avant garde norms. In addition, as described below, choices Parley makes using this listening model can become the basis for talking points. Moreover, we describe below some experiments comparing the music generated before and after editing using this pre-trained model.

A Mood Tagging Model

Parley also employs the *Jamendo MoodTheme* neural model provided in the *Essentia* package (Correya et al. 2021), produced using TensorFlow (Abadi et al. 2015). Given an input audio file, this model has been trained on the MTG-Jamendo dataset (Bogdanov et al. 2019) to output a sequence of vectors, each containing 50 activations (as floats). Each activation refers to a different mood/theme tag and estimates the level of that mood expressed in the audio file. The mood tags include general emotional words such as 'happy', 'sad', 'dark' and 'dramatic'; some associated with musical style or genre, such as 'groovy', 'ballad' or 'christmas'; some associated with settings, such as 'summer' or 'space', and some associated with use-cases for the music, such as 'commercial' or 'trailer'. Sampling an audio file at 32khz, each sample is first passed through a pre-processing model, which encodes it into a latent space, with the resulting latent encoding input to the MoodTheme model. To produce an overall activation vector for an audio file resulting from a passage of music, the activations for each sample can be averaged.

threshold	0.5	1.0	1.5	2.0	2.5	3.0
average tags	8.5	4.8	2.7	1.5	0.8	0.5
zero tags %	4.4	15.9	35.7	54.8	70.5	81.3
ten+ tags %	32.1	12.7	4.5	1.8	0.6	0.2
most used %	28.9	16.0	8.7	5.4	3.1	1.9
tag	'dark'		'space'			'trailer'
least used %	19.5	10.0	5.4	3.0	1.2	0.5
tag	'deep'			'emotional'		

Table 1: For thresholds 0.5 to 3.0, average number of tags per excerpt; % excerpts with no activations for any tag over the threshold and more than ten tags over the threshold; highest and lowest % of excerpts a single tag is assigned to.

To leverage the power of the Jamendo MoodTheme model, we used an early version of Parley to generate 5,000 four-bar excerpts of music recorded in WAV files, then passed each through the model and recorded the average activation vectors produced. Each excerpt comprised a polyphonic melody part and a chord part with three notes which accompanies the melody, as per figure 1. We generated the specifications for Parley randomly, to produce diversity in the generated excerpts, by varying the following:

- The two instruments used for the melody line and chord accompaniment, with each ranging over all 110 non-percussion instruments in the FluidSynth soundfont (see below).
- The volumes (dynamics) of each instrument, ranging over midi volumes 60 to 127.
- The nature of the melody in terms of backbone and passing notes (see below), which changes the number of notes, rhythms and pitch ranges of the notes.
- The tempo of the excerpt, ranging from 1.5 seconds per bar to 3 seconds per bar.

The mean and standard deviations of the activations for the 50 mood tags over the 5,000 excerpts were recorded for use by Parley to produce talking points. In particular, we wanted it to be able to tell when a particular passage of music was exceptionally exhibiting a particular mood. To do this, we looked at the distribution of the model's activations over the 5,000 excerpts, and assigned a mood tag to an excerpt only if the activation it achieved for that tag was T times the standard deviation (for that tag) more than the mean (for the tag), for different thresholds T. The results for thresholds ranging from 0.5 to 3.0 are given in table 1. We see that, if the threshold is set to 0.5 standard deviations above the mean, then excerpts gain 8.5 tags on average, with this decreasing to 0.5 tags for a threshold of 3.0. Moreover, when a threshold of 2.0 is used, more than half the the excerpts are assigned no tag, and the most popular tag (namely 'space') is assigned to only 5.4% of the excerpts, hence no single tag is particularly over-used. Based on these findings, we have implemented in Parley the ability to tell whether a musical passage is *somewhat*, *quite* or *exceptionally* exhibiting a mood if the passage's activation for that mood is 0.5, 1.0 or 2.0 standard deviations above the mean respectively. This means that, referring to table 1, if a passage of music is simi-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

lar to the 5,000 used here, at most 5.4% will be deemed to be exceptionally moody, 16.0% will be quite moody and 28.9% will be somewhat moody, which seems appropriate.

While talking points needn't be perfectly accurate, and users may choose to disagree with them, they must be suitably sensible. Hence we investigated the soundness of the tagging process over the kind of music that Parley can currently generate. To do this, we calculated correlations of activations over the 5,000 excerpts for pairs of tags and inspected the results. We found these pairs to be most negatively correlated, with correlation coefficients less than -0.5:

---

calm & epic; action & calm; action & relaxing; dark & romantic; epic & soft; calm & trailer; action & meditative

---

and we found these pairs to be the most positively correlated, with correlation coefficients above 0.9:

---

dramatic & trailer; epic & trailer; party & sexy; funny & upbeat; groovy & upbeat; melancholic & sad; energetic & fast; action & trailer; dramatic & epic

---

As can be observed, most of the pairs here are reasonable and expected. We did find some unusual results, such as 'sad' and 'happy' having a positive correlation of 0.50 and 'fast' and 'slow' with 0.35. However, in general we found the results sensible and reliable enough for use in Parley.

Neuro-Symbolic Composition

Parley is a modular system with which users can construct a workflow of modules which design, generate, edit and analyse musical compositions, output as both audio files (MP3 and WAV) and as PDF scores which could be played by musicians. Some modules can foreground decisions they make, and analysis modules examine the music and communicate talking points through both text and colour changes to the composition's score. The system is implemented in Python and can be embedded in Colab notebooks (Bisong 2019). This affords a programmatic and a limited GUI interface to Parley, and different notebooks can be used to produce different workflows and hence different compositions. Each notebook first downloads the Parley code repository before using the modules from it.

Cells in a notebook can expose parameters that guide various modules, and then can be used to run those modules, outputting talking point texts, images of the evolving composition's score, midi and audio files. Parley has an internal representation of notes, bars, chords, volume and tempo settings, etc., which can be translated into various formats, including Midi for use in third party music playing apps. It uses the `mido` Midi-handling python package (github.com/mido) and the Fluidsynth (fluidsynth.org) synthesizer to generate WAV files and `ffmpeg` (ffmpeg.org) to turn these into MP3 files. Parley can also output compositions represented as MusicXML (musicxml.com) files, and then employ the `music21` python package (web.mit.edu/music21) and a command-line version of MuseScore (musescore.org) to generate score PDFs. It further uses the `pdf2image` package (pypi.org/project/pdf2image) to turn PDF files into images to display in the Colab notebooks.

Music generated by Parley is episodic, to increase variety, and to enable changes at precise times to accompany other media, e.g,. as a soundtrack to a given video (Colton and Cardinale 2023). The designer, generator and editor modules produce (or change) the represented music for a given episode, with designers producing a skeletal form rather than playable music. Analyser modules produce text describing an episode, communicated in the notebook as talking points. Each module requires a user-supplied *specification* object describing how it should make choices in its processing. There are functions in Parley's codebase to copy and transform specification objects and to copy all specifications for an episode, so that longer compositions can be defined fairly easily.

Each specification object comprises a series of *parameters* which can be set to specify a module's processing. Users can introduce variety and control via the syntax used to represent a parameter. The most straightforward way is to just define the value of a parameter as an integer, float or string. However, parameters can also be defined probabilistically, with a set of values to choose from, each given with a percent probability. For instance, instead of specifying the integer 4 for a parameter, p, the user can specify instead: '3(pc=20) 4(pc=60) 5(pc=20)' for p, indicating that whenever it needs a value for p, it should choose one from $\{3, 4, 5\}$ with probabilities $\{0.2, 0.6, 0.2\}$ respectively. In conjunction with this, users can specify special exceptions, for example in the first (or last) bar of an episode or composition. For instance, p could be specified as '3(pc=20) 4(pc=60) 5(pc=20) 1(cb=1,cb=-1)' with cb standing for (c)omposition (b)ar and following the python standard of -1 indicating the last index of a list. This therefore specifies that p should be set to 1 for the first and last bar of the composition.

Users can also specify cycles for a parameter, for instance '3(ebc=1) 4(ebc=2) 5(ebc=3)' dictates that the value should be 3 in bar one, 4 in bar two and 5 in bar three of every three-(b)ar (c)ycle in the (e)pisode. Finally, users can specify that a parameter should range from one value to another smoothly during an episode. For instance '3(ebp=0.0) 4(ebp=0.8) 5(ebp=1.0)' specifies that the parameter should start when (e)pisode (b)ar (p)roportion is 0.0, rise to 4 when the episode is 0.8 complete, then to 5 when the episode ends.

In the following subsections, each of Parley's modules is detailed with (i) an overview of its purpose (ii) which parameters can be specified and how they affect the process, and (iii) how certain decisions (if any) are foregrounded.

Designers

Parley's designer modules enable users to plan the nature of their composition in advance of generating music for it. In particular, the designers afford description of the structure of episodes (form), the timing of the start and end of bars (tempo) the midi instruments to be used (orchestration) and the chord sequence which will underpin the music.

• The **Form Designer** module takes a string parameter specifying the number of different episodes a composition will have, and the order in which they will play. For instance, the value ABA specifies that the overall composition will comprise two episodes, A and B, with B sandwiched between two A episodes. The two A episodes will not contain exactly the same music, but will be generated by the same modules with the same specifications, thus subject to random variation. After bar timings for each of the episodes have been generated, the form designer constructs *episode objects* to contain bars, and a *form object* to contain the episodes.

• The **Bar Timing Designer** is parameterised with details of (a) the required overall duration (in milliseconds) of the composition (b) the number of episodes required, and (c) the desired duration (in ms) of the bars at the start and the end of each episode. Given these, it calculates the number of bars per episode, and the start and end timings of each bar. It also translates this into ticks for the midi generation (with 960 ticks per second). Details of the calculations for this are given in (Colton and Cardinale 2023). Users must specify the timings (in fractions of bar durations) that chords will change during a bar, using the rhythm format described below.

• The **NRT Chord Sequence Designer** can construct a sequence of trichords (i.e., with three distinct notes in) by employing operators inspired by the music analysis techniques which comprise Neo-Riemannian Theory (NRT) (Cohn 1998). The operators are themselves compounded from smaller units, and this module can be parameterised by the minimum and maximum length of the compounds. Generation of the chord sequences is done by randomly choosing an operator within certain constraints. In particular, the user can specify a *fixed key signature* constraint, so that an operator is only used if it outputs a trichord with all notes in a particular key. Another parameter specifies a *focal pitch* which produces an inversion of the chord such that all the note pitches are as close to this focal point as possible. This module generates a chord sequence to fit the chord change structure specified by the Bar Timing Designer. More details of NRT-based chord sequence generation are given in (Cardinale and Colton 2022) and (Colton and Cardinale 2023).

• The **Orchestration Designer** module is able to take a user-given specification of a mood, such as *happy* or *dramatic*, and produce the details of a pair of midi instruments which increases the likelihood of the generated music evoking that mood. To do this, it uses the specification of the instruments in the file containing 5,000 generative specifications described above. That is, for a given mood tag, T, and a user-specified range, R, this module finds the generative specifications which achieve the top R activations for T, chooses one randomly, and extracts the Fluidsynth soundfont midi instrument number for the melody and chord parts.

Generators

The generator modules in Parley are employed after instruments, episodes, bars and a chord sequence have been designed as above, and adds notes to the bars, sometimes on top of existing notes in a bar. The notes are represented internally with a multitude of properties including pitch, start tick, duration, volume and annotations for the written score, e.g., an indication that a note should be played staccato.

• The **Rhythm Note Sequence Generator** adds a sequence of notes to bars based on a specification which includes details of a rhythm. Such rhythms are described in terms of fractions of a beat for both the start of notes and their duration. They also specify which tone in the relevant chord underlying a bar should be used for the pitch. For instance, the parameter $1 : 1/4 : 1/4, 1 : 3/4 : 1/4$ specifies that in each bar, two notes should play, and each should be the first tone of the chord in the chord sequence at the moment the note plays. The first note starts on the first of four beats in the bar and has duration one quarter of the bar's duration. The second note starts on the third of four beats and also has duration one quarter that of the bar. Users can also specify whether notes should differ from the underlying chord tones by a multiple of 12 semitones (an octave). Normally, multiple note sequence generators are employed to produce chord accompaniments to a melody, as per the case study below.

• The **Voice Leading Melody Generator** adds notes to a bar in three stages. Firstly, guided by the pre-generated chord sequence, it adds chord tones to every bar as the *backbone* of the eventual melody in that bar. The user specifies how many backbone notes per bar, and the sequence is calculated so that no note repeats the one preceding it, if both are mirroring the same chord. When moving to a new chord, repetition is allowed, but the user can specify a *repetition policy* which can disallow/allow this and either (a) simply repeat the note (b) make both notes staccato to introduce a novel rhythm (c) tie the notes into one long one or (d) change the second note to a rest, again to vary the rhythm. This module creates voice leading melodies (i.e., which can be relatively easily sung (Aldwell and Schachter 2010)) by generating only intervals of a tone or a semitone. To minimise the number of pitches between two backbone notes, while adhering to the repetition policy, the backbone sequence is generated to minimise the average difference between the pitches of the notes. In the second stage, passing notes are added between the backbone notes, guided by a user-given *passing notes policy*. This can specify that all or no notes between two backbone notes are added, or only the one note with pitch closest to the midpoint between the backbone pitches is added. The user can specify that the passing notes must all be in a fixed key signature, but if they do not, then the sequence of passing notes changing by a semitone is used to bridge every pair of backbone notes.

In the third stage, the duration of each note is determined using a quantization process. Here, each note is initially given an equal duration of the fraction $\frac{1}{2^n}$ of the bar's duration for n as small as possible. This leaves a shortfall of duration to make up, and the module does this by randomly choosing a backbone note to extend the duration of by another $\frac{1}{2^n}$ until the total duration of the notes adds up to the duration of the bar. For instance, if the backbone and passing notes in a bar comprise 10 notes, then each note is originally given duration $\frac{1}{16}$th of the bar's duration, as this is the largest possible without running over the bar's duration. This produces total note duration of $\frac{10}{16}$ of the bar's duration, so a sequence of 6 backbone notes are randomly chosen (with multiplicity) and extended by $\frac{1}{16}$ to make up the shortfall. Recall that, to increase variety in the music, the specification of these policies can include multiple different choices chosen randomly and/or according to cycles or special cases for particular bars.

• The **Melody Harmonisation Generator** takes a given sequence of notes, usually a melody line, and to each note, n, adds another note with the same starting point in time and the same duration. The pitch of the added note differs to that of n by an interval which is chosen from a user-supplied set of options. The user can specify that the new note pitch must be in a particular key signature, and if so, when none of the intervals achieves this, whether to avoid adding a note or to choose the nearest in-key pitch available. The user can further specify a range of pitches that the new note must be within, and which types of notes to add new notes to, either *all* notes, *passing* notes or *backbone* notes.

Editors

Parley's editor modules take existing compositions and make alterations in the hope of offering improvements.

• The **Note Likelihood Editor** uses the LSTM neural model described above to make predictions about note pitch classes. When it is possible to improve a note's pitch class according to the model, i.e., to be more likely with respect to the training distribution of 18th Century melodies, this is undertaken. Running through all the notes, n, in a user-chosen melody line, the neural model takes as input a sequence of pitch classes from the notes preceding n, and produces a likelihood for the 11 pitch classes available for n. Working at the episode level, the user can specify how many preceding note pitch classes are in the sequence input to the model, with a default of 50 working well. With notes at the start of an episode that do not have 50 preceding notes, the notes at the end of the episode are wrapped around to provide these.

When a pitch class for note n is edited to a new one, the actual pitch must be chosen, and the user can specify how the module should do this. One method involves choosing the appropriate pitch closest to the existing one, while another method chooses the pitch closest to the pitch half way between the preceding and following note pitches (which produces smaller intervals). If a note is given a new pitch class which is more likely than it's existing one, it is marked in green on the score, or blue if the edited pitch class is the same. The user can parameterize this editor with a repetition policy similar to that for the Voice Leading Melody generator, which determines whether an edited note which repeats the pitch of the preceding one, is allowed or disallowed, and if allowed, whether it should be tied, repeated, turned into a rest or made staccato. In practice, when repetitions are not allowed, this means that some notes need to be edited to have a pitch with a lower likelihood than the existing one, in which case they are marked in red on the score.

• The **Orchestration Editor** can take an existing composition and change the midi instruments for individual parts, applied to specific episodes or the composition as a whole, as specified by the user. Alternatively, the user can choose one of the Jamendo MoodTheme tags, and, as per the Orchestration Designer, the module will find pairs of instruments which increase the likelihood that the composition reflects the tag's mood, with the instruments assigned to the parts in the composition, again specified by the user.

Analysers

The analyser modules in Parley serve the purpose of producing talking points that help the user to engage with the music as it is evolving and the processes which are being employed to design, generate and edit the music. Our conception of talking points is evolving along with the development of Parley, and currently includes (a) questions about whether the nature of a passage of music is good enough for the user, which may prompt them to change some aspect of the specifications and re-run a module (b) information about how the process of a module works (c) the identification of certain passages and foregrounding of them by text descriptions and on-score colour changes.

• The **Chord Repetition Analyser** can determine places in a chord sequence where the same chord is repeated, possibly too often. This is subject to a user given window (number of bars) in which to check for repetition, which moves from bar to bar. The module uses the entropy of the chord sequence in a window, and reports the windows with lowest entropy. It combines any overlapping windows into longer ones to report, and marks a user-specified number in red on the score.

• The **Discordancy Analyser** reports any combination of notes in the the composition which are playing at the same time and could be seen as being discordant (which the user may want to avoid for some compositions). The user specifies how long the overlapping notes should play for, in milliseconds, and what intervals should be marked as discordant. The interval of a semitone, i.e., an interval of 1 midi pitch between two notes is usually most noticeable, but an interval of 11 pitches can also be jarring. The user specifies a score for each interval and a total score that must be achieved before a discordancy is reported. The module highlights discordant moments by marking the notes in red.

• The **Likelihood Improvement Analyser** works on compositions, or episodes thereof, on which the LSTM Note Likelihood Editor has been used. It finds those bars where the cumulative improvement (according to the model) is the highest, subject to a user-given threshold of number of edited notes in the bar. Improvement is in terms of total ranking increase from the original pitch class to the edited one, as ranked by the likelihoods returned by the LSTM. This module highlights the chosen bars with purple notes, to contrast the red, green and blue added by the editor.

• The **Mood Highlight Analyser** employs the Jamendo MoodTheme model described above to process the audio from an episode and determine which mood tag is most applicable for it. If the activation of the tag is greater than 0.5 standard deviations (stds) above the mean (over the 5,000 excerpt database), it is reported. The module uses the words *somewhat*, *quite* and *exceptionally* for emphasis if the activation is greater than 0.5, 1.0 or 2.0 stds respectively. The module also finds the passages of b bars in an episode, where the episode's mood tag is most expressed, and reports the k most expressive, with b and k specified by the user.

Case Study: ‘Flaneur’ Compositions

The Parley system currently exists as a codebase which can be downloaded from a repository into a Colab notebook. Each cell in the notebook can set parameters for one or more of Parley’s modules, which can then be run. In practice, a user exposes those parameters of the generative specifications which can fruitfully be experimented with in textbox/slider/checkbox GUI elements, and uses the modules to generate a piece of music in stages, showing audio and score representations of the evolving music at each stage. The same (or other) users can then later vary the parameters and use the notebook to produce novel music.

We used this approach to produce compositions in a series called *Flaneurs*, meaning “connoisseur of the street – a highly observant urban wanderer who takes in everything they see as they seek experiences that fuel their creative minds” (flaneurlife.com/flaneur-meaning). The idea was to juxtapose slower, more melancholy episodes, A, with faster, more upbeat episodes, B, with the music always ambling in nature, constantly in motion. The notebook is available at `https://tinyurl.com/3yjyjtrt`. Sample text and score outputs from the six cells of the notebook are given in fig. 2. The case study can be reproduced and experimented with, using seed 48163 in the Colab notebook.

In cell 1 of the notebook, a *lead sheet* is generated which comprises details of the design of the composition, namely the episode and bar structure and timings, and the chord sequence. The user can specify details for this, with exposed parameters including the overall episodic form, with default ABA given, as well as the composition title, duration in seconds and random seed, if the user wants to recreate a previous session. The parameters for the bar tempos at the start and end of episodes A and B are available to change. Likewise, the chord sequence generation specifications for A and B episodes are also exposed, with default settings for both episode types fixing chords to all be in the key of C major, but with A episodes requiring minor chords (hence fixing chords to Amin, Emin and Dmin) and B episodes requiring major chords (C, F and G), as per the overall idea for the compositions.

In figure 2, we see the cell output showing the episode start times and bars marked, and the chord sequence given. While only the design of the composition, it can still be played audibly, with a piano providing all the parts. Parley has provided one talking point by using the Chord Repetition Analyser to question whether the sequence of four bars of E minor chords is too long. This may prompt the user to change the specifications and re-generate the lead sheet. Note that episode 2 never has two repeating chords in a row, as the `max_repetitions` parameter was set to 1 for the NRT Chord Sequence Generator module of episode 2.

In cell 2, the user specifies the way in which the chords are to be played and the way in which the melody will be initially generated. For the chords, default specifications for usage of Parley’s Rhythm Note Sequence Generator are provided, for example the tonic of the chord is specified to be played with this rhythm: ‘1:3/8:2/8,7/8:2/8(cbc=0,cbc=1,cbc=2) 1:3/8:2/8,5/8:2/8,7/8:2/8(cbc=3) 1:3/8:2/8(cb=-1)’.

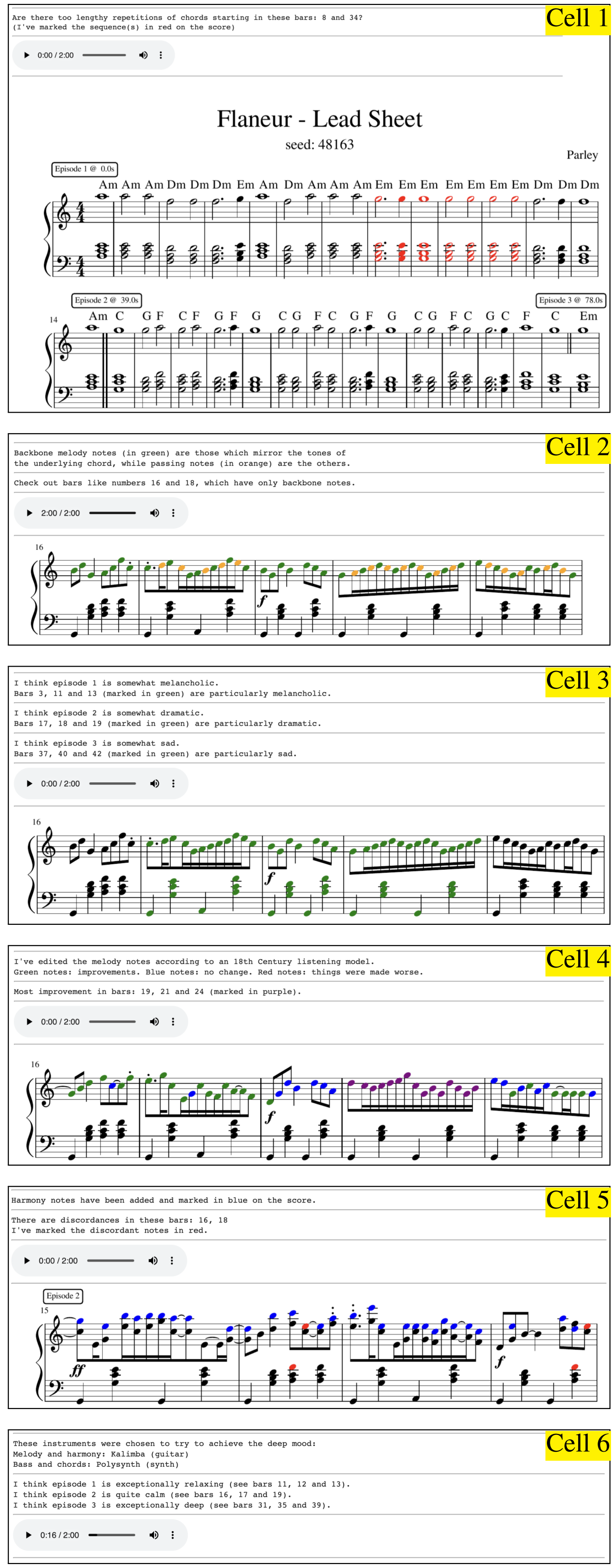


Figure 2: Text and score outputs from the six cells of the Flaneurs Colab notebook for composition with seed 48163. Bars 16, 17 and 18 can be compared as the complexity of the composition evolves.

This states that it should be played on the third and seventh of eight beats for bars 1, 2 and 3 of a 4-bar cycle, but on the third, fifth and seventh of eight beats for bar 4 of the cycle. A special case with (cb=-1) states that for the last bar of the composition, it should play only on the third beat. In every case, the duration is two eighths (one quarter) of the bar duration. Similar specifications for the other chord notes and a bass note are given as defaults in cell 2. The passing notes policy for A episodes is given by default as: 'all(pc=60) mid(pc=30) none(pc=10)', while for B episodes it is: 'mid(pc=50) none(pc=50)'. This ensures a different feel to the two episodes, with larger intervals in B episodes. Parameters for volume changes over the two episode types is also exposed in this cell. The output from cell 2 is shown in figure 2, and we see that the score has green and orange notes highlighted to help the user understand the difference between backbone and passing notes, and Parley points the viewer to bars 16 and 18, where there are no passing notes, as talking points, or at least points of interest.

In cell 3, the user is able to choose instruments instead of the default piano, for all five parts currently in the composition. The generated music is analysed by the Mood Highlighter module and descriptions of the music in terms of mood are provided. As shown in figure 2, the user chose a pan-flute for the melody, keeping the piano accompaniment, and the mood for episodes 1, 2 and 3 were described by Parley as melancholic, dramatic and sad respectively. Note that instead of repeating the most activated tag for multiple episodes, it chooses the second or third as needed.

In cell 4, the Likelihood Improvement editor is used to change the melody line to be more expected in terms of the LSTM's 18th Century music model. The user specified that all notes (backbone and passing) should be edited, and used the repetition policy: 'tie(pc=60) staccato(pc=20) legato(pc=10) disallow(pc=10)', to use when the edited note pitch is the same as the note preceding it. The notes which are changed to having a more likely pitch class (as per the model) are coloured green, while those which stay the same are coloured blue. For instance, we see in figure 2 that in bars 16 and 17, all but one note has been changed. In this cell the Mood Highlight Analyser module is further used to highlight in purple the three 1-bar passages where likelihood improvement is highest, with one such given in fig. 2.

In cell 5, the Melody Harmonisation Generator module is parameterised and employed to add a harmony line to the melody. The default pitch range for the harmony line is given as 60 to 90, so the pitch stays above middle C, the intervals allowed are largely above the melody line (4, 7, 9) with one below (-3), it is told that notes must be in C major, but not to map notes to a key signature, which means that some melody notes are not harmonised. The oboe instrument is used, as its reedy sound compliments the pan-flute. The harmonised notes are highlighted in blue on the score, and the Discordancy Analyser is used to highlight two points where notes differing by 11 semitones are played, as highlighted in figure 2. In the final cell, the user can once again change the instruments used and the volumes for each instrument, in order to produce a final composition. Alternatively, they can choose one of the 50 tags in the Jamendo MoodTheme model, and employ Parley's Orchestration Designer to choose a pair of instruments, one for the melody/harmony and one for the bass/chord. In the case study, the user chose the tag 'deep', and the module chose the instruments Kalimba and Polysynth. The resulting episodes are tagged with talking points 'exceptionally relaxing', 'quite calm' and 'exceptionally deep', with relevant passages highlighted in each episode.

Overall, careful choice of the parameterisations of the generative specifications to produce Flaneur pieces leads to music which is, subjectively, *consistent* due to the repetitive nature of the bass and chords, *varied* (due to the episodic nature, passing note and repetition policy), surprising (due to the likelihood editing), *rich* (due to the harmonisation) and *evocative* (due to the mood-based orchestration). Two 10-minute Flaneur compositions were performed in public in February 2023, as part of a public engagement event at Queen Mary University of London.

Experiments with the Note Likelihood Editor

To investigate the way the likelihood editor changes melodies, we used 100 random seeds to each produce 16 Flaneur compositions via variations on the default generative specifications, and without the harmonisation or instrumentation changes. The melody generation was varied with the passing note policy ranging over 'all', 'mid', 'none' and 'all(pc=33) mid(pc=33) none(pc=34)' which we called 'mixed'. For each policy, we produced four pieces: (i) a voice-leading (VL) version without editing (ii) an edited version (L+) where the process is also constrained to maintain the backbone quality of backbone notes (iii) an edited version (L-) without this constraint, and (iv) a version where a note's pitch class is chosen (R)andomly, but must be in C major. For the L+ and L- versions, pitch classes were also constrained to C major.

We then measured numerous qualities of the generated melodies, in particular certain entropies, with the results given in table 2. The first measurement was the average improvement in the ranking of edited pitch classes over the original, as per the model's calculated likelihood. We see that, for all the L+ and L- versions, the ranking improves by around 5.5, which is half the maximum it could be, of 11. Hence, even with the backbone and C major constraints, the editing is still effective. We next see that the entropy over the pitch classes is high (above 0.9) for all the setups, hence predicting a next note's pitch class is difficult. The same is true for the interval between notes, with the exception of the (unedited) voice-leading compositions with passing notes, where the entropy is around 0.6 to 0.7. Subjectively, these pieces are somewhat dull, as the voice-leading melodies are rather predictable. When edited, the average interval entropies in these compositions rises to 0.9 and above, and – again subjectively – are more varied and surprising. However, the interval entropy for edited melodies is consistently lower than that for the randomly edited melodies, and we have subjectively noted that, while more surprising than pure voice leading, the edited melodies do not seem random.

We also measured the average and maximum *run length* in a composition's melody, with a *run* defined as a sequence

Measurement	All Passing Notes				No Passing Notes				Mid Passing Notes				Mixed Passing Notes			
	VL	L+	L-	R	VL	L+	L-	R	VL	L+	L-	R	VL	L+	L-	R
Edit Improvement Av.	-	5.861	5.842	-	-	5.356	5.337	-	-	5.658	5.644	-	-	5.667	5.644	-
Pitch Class Ent.	0.938	0.939	0.935	0.996	0.962	0.918	0.917	0.990	0.952	0.933	0.934	0.995	0.954	0.935	0.930	0.994
Interval Av.	1.710	2.584	2.367	2.941	3.059	2.388	2.635	2.933	2.062	2.582	2.412	2.939	2.142	2.566	2.426	2.956
Interval Ent.	0.610	0.915	0.895	0.939	0.930	0.856	0.897	0.936	0.650	0.916	0.900	0.939	0.722	0.915	0.898	0.937
Run Length Max	7.000	10.660	23.450	5.980	4.770	7.820	10.970	5.230	6.930	8.490	18.610	5.840	6.900	9.270	19.790	5.650
Run Length Av.	4.078	2.535	2.616	2.481	2.525	2.476	2.524	2.470	3.551	2.512	2.597	2.486	3.413	2.513	2.582	2.472
Run Length Ent.	0.944	0.537	0.501	0.583	0.718	0.536	0.560	0.632	0.835	0.546	0.530	0.599	0.858	0.544	0.515	0.600
Voice Lead Av.	1.000	0.497	0.529	0.419	0.329	0.460	0.421	0.409	0.835	0.505	0.507	0.418	0.788	0.477	0.505	0.413
Voice Lead Ent.	0.000	0.994	0.979	0.978	0.914	0.987	0.955	0.973	0.644	0.986	0.973	0.978	0.736	0.988	0.977	0.977
Backbone Match Av.	1.000	1.000	0.450	0.425	1.000	1.000	0.499	0.424	1.000	1.000	0.452	0.430	1.000	1.000	0.463	0.426
Backbone Match Ent.	0.000	0.000	0.984	0.980	0.000	0.000	0.994	0.981	0.000	0.000	0.986	0.984	0.000	0.000	0.991	0.980

Table 2: Analysis of the properties of melodies averaged over 100 Flaneur compositions, for sixteen different generative setups.

of notes with only positive intervals (i.e., repeatedly going up in pitch), only negative intervals or entirely zero intervals (i.e., repeated notes). To measure the run length entropy, we recorded the lengths of each successive run in a composition's melody and calculated the entropy of this sequence. The results in table 2 help to quantify a phenomenon we observed in the Flaneur melodies after editing, namely long sequences of repeated notes. Recall that the LSTM in the Likelihood Editor listening model requires a seed melody to predict the likelihood of the subsequent pitch classes. As the window only moves by one note, it doesn't change greatly from note to note, hence likelihood calculations can sometimes differ very little over a series of notes, resulting in repeated notes.

The average maximum run length over the 100 L- melodies with all passing notes included was 23.450, which is inflated due to the repetitions. However, the repetitions are broken up with the requirement that backbone notes retain their backbone quality, as the maximum run length reduces to 10.660. Interestingly, while the maximum run length is higher for edited melodies than voice leading ones, the average run length is lower. This again indicates that melodies are less predictable, with many changes in direction and long passages of repeated notes. We also see that the more passing notes in a melody, the more notes in general, hence the higher the potential for repeated notes and increased maximum run length. As run length for edited melodies grows, entropy decreases, which can again be explained by the repetition of notes, rather than increasing or decreasing pitch runs.

In addition to clarifying the editing process, table 2 shows the wide range of entropies for melody properties achievable with certain choices in the generative setup. Hence, there is some evidence that users can control various entropies of the melodies produced, hence can vary melodic expectation (Pearce and Wiggins 2006), and we plan to explore this further. For completeness, we also measured the *voice leading quality* of each composition's melody, giving a score of 1 if a note's pitch is within one whole tone of the previous note's pitch, 0 otherwise. We also measured a *backbone match* calculation, scoring 1 for each backbone note which remained so after editing, 0 otherwise. The results from these evaluations largely conformed to our expectations, as per table 2.

Conclusions and Future Work

We have presented the Parley neuro-symbolic system as a series of generative and analytical modules able to compose music in stages while exposing decisions and opinions as talking points. As with (Aggarwal and Parikh 2020), we believe neuro-symbolic approaches hold much promise for more understandable and interesting generative AI systems. We plan to implement dozens more rule-based modules covering music theory and pre-trained neural models for analysing compositions. As above, the modules will be analogous to composers following rules/heuristics, listening to their evolving compositions and trying to balance global features of the music. We plan to integrate automated reasoning (Robinson and Voronkov 2001) and constraint solving (Rossi, van Beek, and Walsh 2006) techniques to help in the latter tasks.

We also plan to more formally define the notion of a *talking point*. We believe this is a valuable form of framing (Cook et al. 2019; Charnley, Pease, and Colton 2012) creative processes and outputs, which could improve the interactions that users have with a generative AI system, either in terms of casual creation (Compton and Mateas 2015) and entertainment, reflective creation (Kreminski and Mateas 2021) or education (Llano et al. 2020). Once Parley is more sophisticated (including a more intuitive user interface), we plan experiments with amateur and professional musicians and composers to get feedback on the talking points and on Parley in general.

We have already started handing over some creative responsibilities to Parley, for instance with the Orchestration Designer and Editor modules able to choose instruments to (try to) achieve a mood or theme. We are also experimenting with a module which can expand excerpts into entire compositions, following *mood arcs* where the strength of the mood expressed waxes and wanes over the episodic structure of the composition. We hope to investigate the value of neuro-symbolic music composition, where users follow fully (computationally) autonomous processes, with intrigue and explanation provided by talking points about Parley's own compositions, rather than those being made by a user. Ultimately, we aim for Parley to add to musical culture itself, as described in (Colton and Banar 2023).

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments, suggestions and pointers to relevant prior work. We would also like to thank the developers of the various software packages that Parley is built on top of. Berker Banar and Sara Cardinale are supported by the UKRI/EPSRC Centre for Doctoral Training in AI and Music (grant number EP/S022694/1) and Queen Mary University of London.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Aggarwal, G., and Parikh, D. 2020. Neuro-symbolic generative art: A preliminary study. In *Proceedings of the 11th International Conference on Computational Creativity*.

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; Sharifi, M.; Zeghidour, N.; and Frank, C. 2023. MusicLM: Generating music from text. *arXiv 2301.11325*.

Aldwell, E., and Schachter, C. 2010. *Harmony and Voice Leading*. Thomson Schirmer.

Assayag, G.; Rueda, C.; Laurson, M.; Agon, C.; and Delerue, O. 1999. Computer-assisted composition at IRCAM: From PatchWork to OpenMusic. *Computer Music Journal* 23(3):59–72.

Banar, B., and Colton, S. 2022. Identifying critical decision points in musical compositions using machine learning. In *Proceedings of the 24th IEEE International Workshop on Multimedia Signal Processing*.

Bisong, E. 2019. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress.

Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019. The MTG-Jamendo dataset for automatic music tagging. In *Proceedings of the ML4MD Machine Learning for Music Discovery Workshop at ICML*.

Cardinale, S., and Colton, S. 2022. Neo-Riemannian Theory for generative film and videogame music. In *Proceedings of the 13th International Conference on Computational Creativity*.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the notion of framing in computational creativity. In *Proceedings of the 3rd International Conference on Computational Creativity*.

Cohn, R. 1998. Introduction to Neo-Riemannian Theory: a survey and a historical perspective. *Journal of Music Theory* 42:167–180.

Colton, S., and Banar, B. 2023. Automatically adding to artistic cultures. In *Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art and Design*.

Colton, S., and Cardinale, S. 2023. Extending generative Neo-Riemannian Theory for event-based soundtrack production. In *Proceedings of the International Conference on Computational Intelligence in Music, Sound, Art and Design*.

Compton, K., and Mateas, M. 2015. Casual creators. In *Proceedings of the 6th International Conference on Computational Creativity*.

Cook, M.; Colton, S.; Pease, A.; and Llano, T. 2019. Framing in computational creativity: A survey and taxonomy. In *Proceedings of the tenth International Conference on Computational Creativity*.

Correya, A.; Alonso-Jiménez, P.; Marcos-Fernàndez, J.; Serra, X.; and Bogdanov, D. 2021. Essentia TensorFlow models for audio and music processing on the web. In *Proceedings of the Web Audio Conference*.

Hitzler, P.; Eberhart, A.; Ebrahimi, M.; Sarker, M. K.; and Zhou, L. 2022. Neuro-symbolic approaches in artificial intelligence. *National Science Review* 9(6).

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Kreminski, M., and Mateas, M. 2021. Reflective creators. In *Proceedings of the 12th International Conference on Computational Creativity*.

Llano, M.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable computational creativity. In *Proceedings of the 11th International Conference on Computational Creativity*.

Pearce, M., and Wiggins, G. 2006. Expectation in melody: The influence of context and learning. *Music Perception* 23:377–405.

Robinson, A., and Voronkov, A., eds. 2001. *Handbook of Automated Reasoning*. MIT Press.

Rossi, F.; van Beek, P.; and Walsh, T., eds. 2006. *Handbook of Constraint Programming*. Elsevier.

5. Climate Change, Diversity, Equity, and Inclusion

What Does Genius Look Like?
An Analysis of Brilliance Bias in Text-to-Image Models

Juliana Shihadeh and Margareta Ackerman
Department of Computer Science and Engineering
Santa Clara University, Santa Clara, California
{jshihadeh,mackerman}@scu.edu

Abstract

As text-to-image models and the visuals that they create become increasingly integrated into society, it is imperative to develop an awareness of the inherent biases within these technologies. While earlier visual creative machines such as AARON by Harold Cohen (Cohen 1999) and The Painting Fool by Simon Colton (Colton et al. 2015) have exhibited remarkable creativity, the methodology underlying today's popular text-to-image models rely heavily on public data to produce visuals, resulting in an increased risk for bias. Further, recent image generation technologies, such as Dall-E (Q.ai 2022) and Midjourney (Salkowitz 2022) and applications such as LensaAI have attracted millions of users (Curry 2023), making it more urgent to ascertain the risks of these technologies. In this paper, we initiate an analysis of text-to-image models focusing on Brilliance Bias, a negative stereotype of women's intellectual abilities and holds back women's potential. Our findings reveal a significant presence of Brilliance Bias in Dall-E, Midjourney, and Stable Diffusion.

Figure 1: Sample output of Midjourney prompted on "Genius person." The parameters are set to generate four images per output.

"You can't be what you can't see"
-Lean In Organization and Getty Images[1]

Introduction

Creative machines have long been made and studied within academia. When it comes to machines creating art, one of the earliest examples includes Harold Cohen's AARON, which Cohen taught to draw and later paint in his own style (Cohen 1999; Sundararajan 2021). AARON has been showcased in galleries as early us 1995 (Garcia 2016). Another notable system is The Painting Fool by Simon Colton, which unlike AARON, aimed to be taken seriously as an artist in its own right (Colton et al. 2015). The Painting Fool and its work have been showcased at public venues, such as the 2013 Paris exhibit "You can't know my mind" (Shubber 2013). Industry involvement in the arena of creative machines was gradual, with systems such as Google's DeepDream entering the scene in 2015 (Rayner 2016).

Microsoft's investment in OpenAI began to change the landscape, focusing on the creation of large (and expensive) models at a magnitude that was not previously possible with respect to amounts of data used for training and the size of the models. The introduction of the text-to-image model DALL-E, leading to Stable Diffusion and proliferation of commercial apps, such as LensaAI, brought generative AI visuals to the masses. AI-generated art is now incorporated into advertising (Nestle (Kiefer 2022), future Super Bowl ads (CBInsights 2023), and Rosebud.ai (Koidan 2020)). At the same time, firms such as Microsoft (Microsoft 2023; Q.ai 2022), Canva (Adams 2022), and Shutterstock (Shutterstock 2023) have integrated image generative capabilities into their products.

Text-to-image generation through large models does not come without pitfalls. One of the main concerns with these models is their reflection, and perhaps even amplification, of biases present in the data they are trained on. Researchers study racial bias (Agarwal et al. 2021; Wiggers 2021; Srinivasan and Uchino 2021a) in these models, and gender bias is analyzed with respect to clothing and physique in text-to-image models associations to women and men (Chiriguayo and Ta 2022; Steele 2022).

A little known, but significantly impactful, bias called

[1]leanin.org/getty, Accessed: 2/27/2023

Figure 2: “Genius person” prompted to Dall-E. The top and bottom rows show examples of women and men respectively.

Figure 3: “Brilliant person” prompted to Dall-E. The top and bottom rows show examples of women and men respectively.

“Brilliance Bias,” hinders some of the most high-potential members of our society. Brilliance Bias is the association of higher intellectual capabilities to males (Leslie et al. 2015), that is, the implicit belief that intellectual brilliance is more likely to be present in men than in women. It impedes women’s potential through both their self-perception and the opportunities that others are willing to grant them.

We initiate the study of Brilliance Bias in large text-to-image models. Our analysis focuses on evaluating its presence in some of the most popular models, namely Dall-E, Stable Diffusion, Midjourney and Craiyon (formerly Dall-E mini). Visuals influence people’s perception of the world. For example, a review on stock photos showing stereotypical depictions of women such as in supporting roles proves to negatively impact women’s career potentials (Miller 2014). A developmental psychology study on media from 2000-2020 reveals that media significantly influences young people’s views on gender roles (Ward and Grower 2020).

Given the rapidly growing popularity of text-to-image models and the powerful societal influences of images, it is critical to understand the biases exhibited by these models. As Sun et al. (Sun et al. 2022) points out, “social transparency – making visible the socio-organizational factors that govern the use of AI – can help users form a socially situated understanding of an AI system and take more effective actions with it.” A clear insight into the presence of

Figure 4: “Brainiac person” prompted to Dall-E. The top and bottom rows show examples of women and men respectively.

the biases found in text-to-image models will help find effective solutions to mitigate those biases, and limit their impact. Our study initiates an analysis of the Brilliance Bias in these models, acting as an essential first step to mitigate the amplified impact of this bias.

Background

Brilliance Bias

Despite the numerous intellectual contributions made by women, their intellectual abilities are consistently downplayed through a pervasive bias known as “Brilliance Bias.” Brilliance Bias is the implicit belief that intellectual brilliance is a male trait. This bias is found to be pervasive in the STEM and Humanities fields, and correlates to lower female to male ratios of PhD students studying Computer Science, Mathematics, Philosophy and Music Composition (Leslie et al. 2015). Studies on children show that it starts as early as 5-7 years old, as seen by children selecting boys in a game for “really, really smart” teammates (Bian, Leslie, and Cimpian 2018). When asked to pick out images associated with stories and descriptions of being “really, really smart”, girls are less likely to pick from their gender and more likely to associate to being “really, really nice” starting at the age of 6 (Bian, Leslie, and Cimpian 2017). Furthermore, at the age of 6, girls’ interests shift because they think of themselves as less brilliant; they are more likely to pick a game for children who try “really, really hard” and less likely to pick games for “really, really smart” children. At the age of 5, “really, really smart” children’s games are more equally selected by boys and girls.

Research focusing on STEM fields shows that both women and men affiliate brilliance to STEM (Deiglmayr, Stern, and Schubert 2019) (the belief that people who are in STEM are brilliant). Furthermore, the study shows men are less likely than women to believe in the existence of Brilliance Bias and more likely to feel like they belong in STEM fields. It is further shown that women are less likely to be referred to jobs that require high levels of intellectual ability (Bian, Leslie, and Cimpian 2018).

Brilliance Bias has only recently began to be studied in the context of generative models. Last year, an adjective

Figure 5: "Brainiac person" prompted to Midjourney. Of 100 images, only two are identified as female. They are shown in the top row, left to right. The rest of the images are identified as male.

and lexicon study on Brilliance Bias in large text models, specifically OpenAI's models, reveals a significant presence of Brilliance Bias. When the OpenAI models are prompted with identical brilliance prompts other than gender, men are associated with higher levels of power, agency, valence, arousal, and dominance (Shihadeh et al. 2022).

Biases in Images

A Google Search Engine study analyzes the occupational gender biases in image search queries (Kay, Matuszek, and Munson 2015). Its results show a significant representation of stereotypical gender roles and minorities, such as women, portrayed unprofessionally in images. Furthermore, it points out people are more likely to use image results that align with their stereotypical beliefs causing a dangerous loop of increasing biases.

One paper looks at the bias of CEO genders in the Google Search Engine and finds that results are dominated with white men (Lam et al. 2018). Another study finds that even though efforts were put to mitigate the gender bias of the query "CEO", combinations of "CEO" with a country such as "United States" resurface the gender bias (Feng and Shah 2022). Thus revealing the challenges in fully mitigating a bias that is deeply embedded in a system.

Studies on facial recognition show a bias in being able to identify white men more accurately, in particular significantly misclassifying black women as male (Raji et al. 2020). An analysis of image recognition models shows that images of women are annotated more on appearance and less likely to be identified in image detection technology compared to men (Schwemmer et al. 2020). If image recognition tools are used to annotate and label images for training text-to-image models, computer labeling biases could further increase gender biases society gets exposed to.

Biases in Generative AI

While biases are studied in text-to-image models, no prior research of this kind focuses on Brilliance Bias. For example, gender bias in occupations is found in the text-to-image model CLIP (Wiggers 2021; Agarwal et al. 2021). A high correlation of stereotypical occupations is found associated

Figure 6: "Genius person" prompted to Craiyon. No images out of the 100 images we generate display a woman.

Figure 7: "Brilliant person" prompted to Craiyon. The top and bottom rows show examples of women and men respectively.

to women, such as "nanny" and "housekeeper", and men, such as "prisoner" and "mobster". Furthermore, racial biases are found such as black people misclassified to be non-human, being labeled as "animal", "gorilla", and "chimpanzee". Additional racial biases are found on lightening the skin tone of a person (Srinivasan and Uchino 2021a; Mattei 2022). One study finds race and gender biases in Stable-Diffusion with descriptive phrases like "emotional" showing women and "poor" showing more dark skinned people (Bianchi et al. 2022). A study on cycleGAN examines how an art style miscaptured in generative models can cause "inaccurate information about socio-political-cultural aspects" (Srinivasan and Uchino 2021b).

Generative AI app users note seeing their race being erased (Mello-Klein 2022; Sung 2022). Others point out Asian women in particular being depicted in tears and showing more nudity (Heikkilä 2022). Some users see stereotype portrayals of women, such as slimming waists (Chiriguayo and Ta 2022) furthermore exposing women's skin and anatomy more, while men are more likely shown in professional apparel (Steele 2022; Heikkilä 2022). OpenAI attempted to add more diversity to DALL-E, particularly as it applies to occupation (OpenAI 2022), and maybe appending words like "black" and "female"[2]

[2]https://twitter.com/minimaxir/status/1549070583035416576

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Figure 8: "Brainiac person" prompted to Craiyon. The top and bottom rows show examples of women and men respectively.

How Visuals Affect Society

Images are an integral part of our world. Research that looks at how images affect students' learning in middle school concludes that images influence their understanding of the world, finding that "if you look at an image, it puts more ideas in your head" (Hibbing and Rankin-Erickson 2003). Furthermore, based on the cultivation theory, repeated exposure over time alters one's perception of the world (Potter 1993; Shrum 1995). One study finds that short term exposure also affects one's views. It finds that skewing Google search results changes people's choice in selecting a woman or man to represent a job (Kay, Matuszek, and Munson 2015). Another study finds that stock photos put women in supporting roles, stereotyped roles, and sexualized their images further finding that seeing these images hurts women's career aspirations (Kay, Matuszek, and Munson 2015; Suddath 2014). This work, led by Sheryl Sandberg and Getty Images, resulted in the initiative of "You can't be what you can't see" (LeanIn.Org 2023). To mitigate visual biases, they curate a set of creative images with archetypes rather than stereotypes; these images portray diverse examples of families, women in powerful roles and men as caretakers in addition to earners.

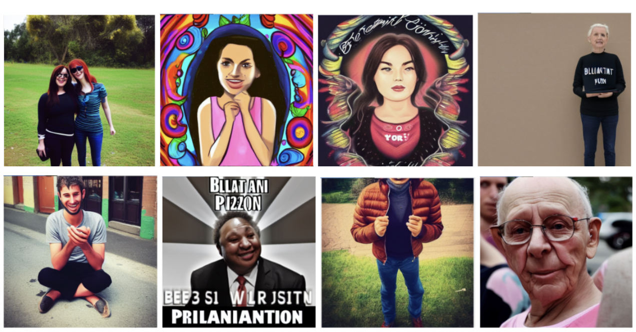


Figure 9: "Brilliant person" prompted to Stable Diffusion. The top and bottom rows show examples of women and men respectively.

Multiple studies on media reveal it has a powerful influence on society. Stereotypes influence a person's inclination to join a field, changing how media, such as television for instance, portrays computer scientists can in turn help with demonstrating the diversity of a field (Cheryan et al. 2013). Due to the "digital generation", teens are especially prone to being influenced by media about how they see themselves and socialize (Celestin 2011). For instance, Silicon Valley and the Big Bang Theory show women as a background character, usually for the role of a love plot in a story rather than a leader (Javed 2015). Furthermore, these TV shows have a stereotypical nerd association to the male characters which can be discouraging for girls' perception of a field (Javed 2015; Welsh 2013). Supporting studies show that girls who see stereotypical portrayals or behaviors of people are more likely to demonstrate the stereotypical behavior themselves (Essig 2018). Geena Davis who is an advocate of more women in film leadership roles, shows the film industry influences women's ambitions, changes toxic relationship dynamics, and encourages success (Ford 2019; Institute 2016). The mass effect of media on people's perceptions of the world and themselves demonstrates how influencing visuals are.

Methodology

We analyze the output of four text-to-image models to determine whether these models exhibit Brilliance Bias. The models we evaluate are Dall-E, Midjourney, Craiyon and Stable Diffusion. To study the presence of this bias, we provide each model with a set of brilliance prompts, designed to elicit the creation of an image of a person the model deems "genius" or "brilliant." Furthermore, we test the models on the base case prompt "person" to compare against brilliance prompts. To analyze the results, we evaluate the differences of the number of women and men in the generated output.

Data

Figure 10: "Super Smart person" prompted to Stable Diffusion. The top and bottom rows show examples of women and men respectively.

We generate 400 images using each text-to-image model for four different brilliance traits[3]. Instead of feeding the models a single prompt, like "Brilliant person," we expand

[3]Our data can be found at https://github.com/julishi/Brilliance-Bias-in-Text-to-Image-Models/tree/main

our analysis to a set of carefully designed prompts. The reason is that words inherently have multiple meanings, and if we seek to understand how models visualize intellectual brilliance, it is best to tackle this challenge through several prompts that all aim to uncover this aspect of the models. The brilliance traits (other ways to say "brilliant") that we use are based on the ones selected in Storage et al.'s (Storage et al. 2020) study to analyze if people associate brilliance with men more than women. These are "brilliant", "genius", "brainiac", and "super smart."

Each prompt is constructed as "[trait] person," resulting in the following 4 prompts: "Brilliant person", "Genius person", "Brainiac person", and "Super Smart person." We capitalize all the traits. We use the word "person" with each trait to neutralize the gender of the trait in our prompt and guide the models toward creating a human. The aim is to determine whether the model tends to identify high intellect with men or women.

To more accurately ascertain the models' Brilliance Bias versus other forms of gender bias, we have decided to test the models' behaviour on the more basic prompt "person." This exploration is motivated by the presence of such a bias in humans, whereby people assume that gender neutral words refer to men (Bailey, Williams, and Cimpian 2022). We label it as "None" in our graph results.

We generate 100 images per prompt, totalling to 500 images per model. All together, we look at 2000 images across Dall-E, Midjourney, Craiyon and Stable Diffusion. In this set of experiments, we intentionally avoid specifying style in order to reduce the risk of additional influences.

We run: Dall-E on its website, Midjourney on its discord, Stable Diffusion on its DiffusionBee app, and Craiyon via its website too. We run each prompt on Dall-E and Midjourney 25 times, with each generating 4 images to create 100 images. Craiyon produces 9 images per prompt, so we run each prompt 12 times and take the first 100 images. We set Stable Diffusion to generate 100 images per prompt and set its Guidance Scale to the maximum 20 in order to understand its behavior when it is more strongly influenced by the prompt.

Across the gender spectrum, for the purposes of our study we focus on Brilliance Bias in the context of the binary genders male and female. We use terms representing binary genders such as "male" and "female" and "woman" and "man" in our paper as shorthand for a figure identified by our analysis as exhibiting binary male-identifying or female-identifying traits.

Once all the images are generated, we look at how many are of a woman vs a man to study if high intellect is more often associated with men or women. In order to do this, we manually count the number of women and men in these images. We count our images based on 3 categories: Male, Female, and Other. For an image we could not determine a gender or that did not have a person, we count it as "Other." Although rare, some images display multiple people (most often seen in Craiyon and Stable Diffusion). If an image includes at least one male and at least one female, we label it as "Other." Most images portray a single person, and are easily classified as showing male-identifying or female-identifying traits looking at a combination of physique, clothing and facial features. We expect negligible deviation if multiple people were to label the images. [4]

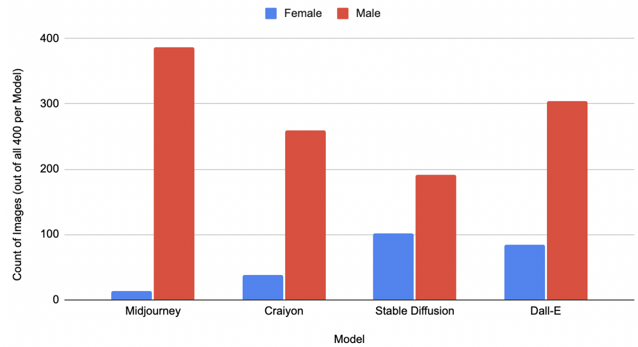


Figure 11: **Comparing models' net count of female and male images for the brilliance prompts**. We created 400 images per model, 100 images for each brilliance trait we evaluated: "genius", "brilliant", "super smart" and "brainiac".

Results

Our analysis demonstrates a clear presence of Brilliance Bias in Midjourney, Stable Diffusion, and Dall-E. The results for Craiyon are inconclusive due to gender bias seen on "person" as well. Furthermore while Stable Diffusion clearly demonstrates Brilliance Bias, it is less biased than the other models because of its performance on the prompts "Brilliant person" and "Super Smart person".

We consider the overall ratio of generated images of women to images of men across all the traits we test for each model, shown in Figure 11. In most cases, the models produce at least twice as many men as women on the brilliance prompts, often with the disparity being much greater. Midjourney shows the greatest disparity in number of images of women to men, with 3.25% women vs. 96.5% men, followed by Craiyon 9.5% women vs. 65% men, Dall-E 21% women vs. 76% men and Stable Diffusion 25.25% women vs. 47.75% men. The only exception is Stable Diffusion, which shows a ratio slightly below 2x. These results are rather unfortunate, since Midjourney is known to incorporate more art style.[5]

[4]For completeness, we look at studies of gender assigning based on facial features and clothing. Men are found to have a more prominent chin/jaw and protuberant nose/brows (Bruce et al. 1993). Women are found to have higher eyebrows while men have thicker eyebrow closer to their eyes (Brown and Perrett 1993). Women are noted to have fuller cheeks and less facial hair including around their eyebrows, while men have more facial hair or hair follicles otherwise (Burton, Bruce, and Dench 1993). A study looking at the Halloween clothing of children found female clothing are more decorative and exposing of skin, while male clothing are more functional (Murnen et al. 2016).

[5]https://simplified.com/blog/ai-text-to-image/dall-e-2-vs-midjourney/, https://startuptalky.com/dall-e-vs-midjourney/

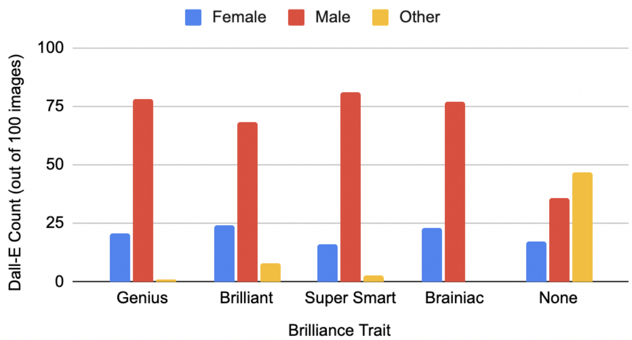


Figure 12: **Dall-E** Brilliance Bias results. The number of Female, Male, and Other count for each Brilliance Trait tested.

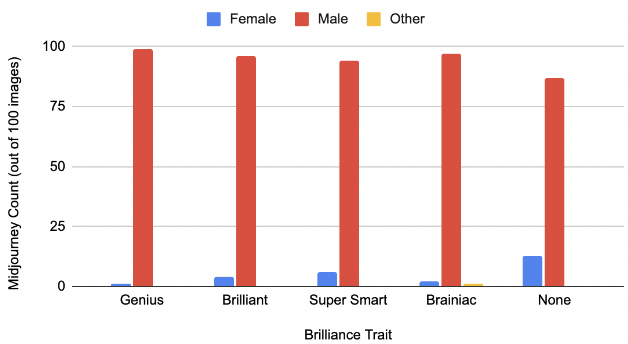


Figure 13: **Midjourney** Brilliance Bias results. The number of Female, Male, and Other count for each Brilliance Trait tested.

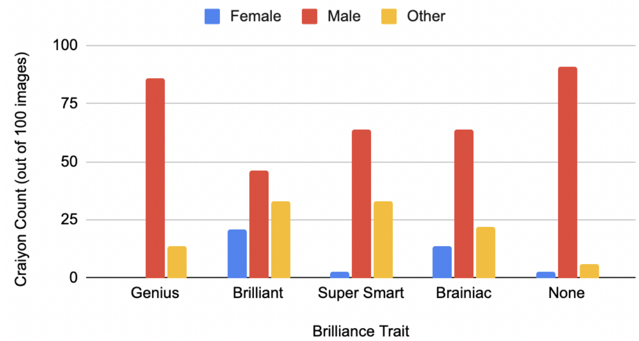


Figure 14: **Craiyon** Brilliance Bias results. The number of Female, Male, and Other count for each Brilliance Trait tested.

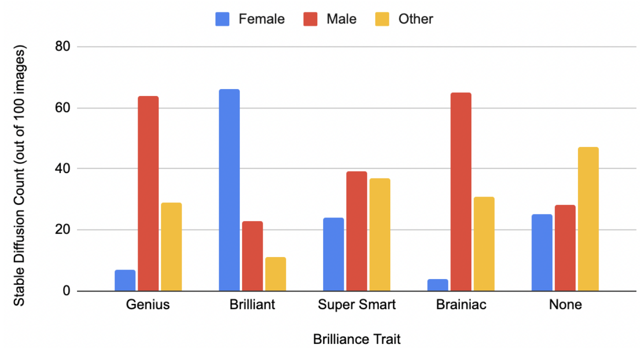


Figure 15: **Stable Diffusion** Brilliance Bias results. The number of Female, Male, and Other count for each Brilliance Trait tested.

Across all four models, nearly all prompts result in a significantly stronger association of high levels of intellect to men. Midjourney in particular has the largest gap between men and women, as shown in Figure 11. Craiyon and Stable Diffusion, seen in Figures 14 and 15, have the highest number of images labeled as "Other" amongst the models studied. Meanwhile, Stable Diffusion is an exception on the prompt "Brilliant person" resulting in a higher female:male ratio as seen in Figure 15 compared to the other models. This could be due an alternate meaning of "brilliance," which can be defined as "full of light, shining, or bright in color".[6] Stable Diffusion's images on "Super Smart person" results in the closest count of images between all three categories: Female, Male, and Other.

Midjourney produces almost no women for the prompts "Genius person" and "Brainiac person." Craiyon generates no women out of 100 images for the prompt "Genius" and almost no women for the prompt "Super Smart person." There is no one consistent brilliance trait that shows the highest level of Brilliance Bias across all the models, as each model varies in performance on the four traits. However, all the models show a significant difference between the number of female and male images for brilliance prompts.

We compare how the ratios of male to female images in brilliance prompts contrast to the same ratios on the prompt "person." Stable Diffusion has the closest 50:50 female:male ratio on the non-brilliance prompt "person" as seen in Figure 15. On the other hand, the rest of the models have a more notable higher count of men to women when prompted to generate a "person". However, Midjourney's results in Figure 13 show it generates less women for brilliance invoking prompts compared to the non-brilliance prompt "person". For these three models, brilliance prompts lead to a much greater difference in the number of women vs men generated, suggesting strong evidence of Brilliance Bias.

On the contrary, Craiyon generates a higher count of men for both the brilliance prompts and non-brilliance prompt as seen in Figure 14. This makes it more challenging to separate Brilliance Bias from other forms of gender bias for this model. Future work will be needed to assess the "default male" bias and separate it from Brilliance Bias in the Craiyon model.

Discussion

A comparison of the models' performance on brilliance prompts and the non-brilliant prompt "person" indicates Stable Diffusion, Midjourney, and Dall-E are Brilliance Biased

[6]https://dictionary.cambridge.org/us/dictionary/english/brilliant

while Craiyon's Brilliance Bias is questionable for the time being. Since these models are trained on data created by people, they are simply revealing the biases that exist in society. The Brilliance Bias we are seeing here is a mirror of the collective unconsciousness of society at large. However, these models, and biases they embody, will influence society on a large scale due to their popularity and the influence of images and media on people. Consequently, they will stand to hold back inclusivity progress. Below we discuss a comparison of how the models' portray brilliant men and women. Furthermore, we discuss how we can mitigate Brilliance Bias in text-to-image models.

How Brilliant Men and Women are Portrayed

We observe notable differences in how text-to-image models visualize intellectually brilliant men vs women. One such difference came across in the prompt "Brilliant person." In this case, we find that Dall-E visualizes the "brilliance" emanating from the men, while for female characters, the brilliance is visualized as a decorative environmental factor. See Figure 3.

The results also suggest that the term "brilliant" more often represents the non-intellectual interpretation of the term when it came to women, "full of light, shining, or bright in color." [7] This came across in higher adornment of women with fancy jewels, makeup, and radiant smiles, which was not the case for the generated images of men under the same prompt, whose visualization better align with the intellectual interpretation of the term "brilliant." See Figure 3.

Stable Diffusion makes images of brilliant men more often photorealistic, while brilliant women are visualized in a more artistic fashion, as seen in Figure 9. It is interesting to see that Stable Diffusion shows groups of women multiple times when prompted with "Brilliant person," as well, compared to more often showing a man by himself. This appears to imply that women are not individually capable of holding high-levels of intellect, rather it is through a group effort that they achieve brilliance. This may reflect unconscious biases in society, absorbed through the images that the models are trained on.

Additionally, we notice objects around women's heads more often compared to men, for example, as seen in Figure 2 with a light bulb and cloud above two women's heads. However, such illustrative elements are not as commonly seen above men's head. Why do the models end up adding these objects for brilliant women but not men? This seems to suggest that to appear convincingly brilliant, a woman needs visualizations of her thinking, while a man's intellect can be assumed without such props. In future work, it may be interesting to analyze the items (ex. swirling icons, thought bubbles, items emanating from a person's head) that tend to co-occur in generated visualizations of brilliant men vs women.

Moreover, we notice multiple images cut off women's faces in Dall-E and Stable Diffusion. This can be seen in Dall-E's images in Figures 2 and 3, and in Stable Diffusion's Figure 10 top row last image. Even more so, we find images Craiyon creates, in particular for the prompt "Brilliant person" and "Brainiac person", portray women with more exposed skin, as seen in Figures 7 and 8, and nudity. Furthermore, Midjourney more often shows men as cyberborgs as seen in Figure 5. For the "genius" prompt, Craiyon generates zero women as seen in Figure 6.

The above summarizes our observations. Further analysis would be needed to conclusively report on the above.

General Stylistic Elements

Across all the models, we note a few generic stylistic elements. For instance "brainiac" is affiliated with green colors, robotic-like figures, and persons that have a Frankenstein-like look too. These images resemble the comic book character "brainiac"[8], potentially suggesting that for this prompt the models may be more influenced by that character than the intellect-related definition of "brainiac." Stable Diffusion incorporates more colors to "brainiac" though, particularly pink and purple In addition, we notice Dall-E often times shows a brain with "brainiac" as seen in the images in Figure 4. Furthermore, the trait "super smart" results in common superman stylistic details across all models, including caps and red and blue colors. Additionally, Midjourney shows a person's face the most clearly but adds some artistic texture, with Dall-E showing a person in a photographic style more often. Craiyon least often shows a real-person. Stable-Diffusion most often adds text to images, although a majority of the time it did not make sense. Lastly, Midjourney affiliates "brilliant" and "genius" less often to younger people compared to the other models.

Mitigating Brilliance Bias

We explore purposefully altering the style specified in a prompt to see if it can help mitigate the Brilliance Bias we found. We assume adding the keyword "contemporary art style" might influence the models to generate more gender inclusive images. This is in consideration that society has progressed (to a certain degree) toward being more inclusive of women and thus we hypothesize that a contemporary style would reflect that. However, an exploratory analysis shows that just adding "contemporary art style" keeps most of the images male-dominant.

We further explore the specific contemporary art style "Feminist art", defined as a "movement [that] arose in an attempt to transform stereotypes and break the model of a male-dominated art history" (Invaluable 2021), and find it to very clearly increase the number of images that had a woman or women. This is not too surprising though given that the art style focused on enhancing the representation of women, making a point that the text-to-image models are representing society's cultures and beliefs accurately. However, it is worth exploring the variety of art styles more in depth in future work. Furthermore, the word "feminist" itself tends to be associated with women, and may prompt more images of women as these models often use words that appear in the prompts out of the context.

[7] https://dictionary.cambridge.org/us/dictionary/english/brilliant

[8] https://www.dc.com/characters/brainiac

Conclusions and Future Work

In this paper, we evaluate the presence of Brilliance Bias in four text-to-image models: Dall-E, Midjourney, Craiyon and Stable Diffusion. Our results reveal that text-to-image models show men much more often than women when asked to generate a person portraying brilliance.

There is a substantial presence of the Brilliance Bias in the Dall-E, Midjourney, and Stable Diffusion. The results are more ambiguous in the case of Craiyon as it reveals gender bias regardless of brilliance. Midjourney and Stable Diffusion generate fewer images of women on brilliance traits compared to the non-brilliant prompt "person." Dall-E often presents gender-neutral images when prompted to create a "person", while associating brilliance to men. Midjourney shows the most significant difference in ratio of women compared to men when given brilliance prompts, with women shown in only 3.25% of its images. Craiyon created 9.5% images of brilliant women, followed by Dall-E with 21%, and Stable Diffusion with 25.25%.

This analysis leads us to realize that there is another fundamental bias that needs to be studied in text-to-image models. That bias, which has been found in humans, is the tendency to assume that gender neutral terms such as "person" refer to men rather than women (Bailey, Williams, and Cimpian 2022). Craiyon generates more images of men than women for brilliance induced prompts. However, it creates even more images of men when prompted with the non-brilliance prompt "person." Thus, Craiyon seems to exhibit a more fundamental bias, the assumption that people are men, making it more challenging to ascertain the extent to which it exhibits Brilliance Bias.

We hope that this work spurs interest in further analysis as well as mitigation of biases in generative models, particularly those that are widely accessible. We have conducted an initial analysis into this foray. In particular, the bias whereby the models assume that general neutral words refer to men deserves further study. One of the greatest challenges arising from our results is the mitigation of Brilliance Bias in generative models. Solutions can come in the form of creating new models that do not exhibit this bias, or corrective tools that work in conjunction with large models. While in this initial study we focus on male vs female analysis of Brilliance Bias, it is worth expanding this analysis across the gender spectrum.

Images play a critical role in influencing people's perception of themselves, their abilities and of the potential they see in themselves and others. Given text-to-image models are rapidly growing in popularity, it is important to understand their biases to help mitigate their spread. Rather than introduce biases that set back progress society makes on inclusivity efforts, it is important to navigate these popular image generators toward a more equitable and diverse representation of society.

References

Adams, C. 2022. Turn imagination into reality with text to image in Canva. https://www.canva.com/newsroom/news/text-to-image-ai-image-generator/. (accessed: 02.27.2023).

Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating clip: towards characterization of broader capabilities and downstream implications.

Bailey, A. H.; Williams, A.; and Cimpian, A. 2022. Based on billions of words on the internet, people = men. *Science Advances* 8(13):eabm2463.

Bian, L.; Leslie, S.-J.; and Cimpian, A. 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science* 355(6323):389–391.

Bian, L.; Leslie, S.-J.; and Cimpian, A. 2018. Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist* 73(9):1139.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*.

Brown, E., and Perrett, D. I. 1993. What gives a face its gender? *Perception*.

Bruce, V.; Burton, A. M.; Hanna, E.; Healey, P.; Mason, O.; Coombes, A.; Fright, R.; and Linney, A. 1993. Sex discrimination: how do we tell the difference between male and female faces? *perception*.

Burton, A. M.; Bruce, V.; and Dench, N. 1993. What's the difference between men and women? Evidence from facial measurement. *Perception*.

CBInsights. 2023. The future of the Super Bowl Ad: How tech like generative AI is automating TV ad creation production. https://www.cbinsights.com/research/future-of-super-bowl-tv-advertising/. (accessed: 02.27.2023).

Celestin, M. 2011. Empowering and engaging teen girls through media from the perspective of a practitioner and producer.

Cheryan, S.; Plaut, V. C.; Handron, C.; and Hudson, L. 2013. The stereotypical computer scientist: Gendered media representations as a barrier to inclusion for women. *Sex roles*.

Chiriguayo, D., and Ta, A. 2022. Lensa AI portrait app may be using old beauty standards and male biases. https://www.kcrw.com/news/shows/press-play-with-madeleine-brand/ai-misogyny-migrants-soccer/lensa-ai. (accessed: 02.27.2023).

Cohen, H. 1999. Colouring without seeing: a problem in machine creativity. *AISB quarterly* 102:26–35.

Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M.; and Ferrer, B. P. 2015. The Painting Fool sees! new projects with the automated painter. In *ICCC*, 189–196.

Curry, D. 2023. Lensa AI revenue and usage statistics. https://www.businessofapps.com/data/lensa-ai-statistics/. (accessed: 02.27.2023).

Deiglmayr, A.; Stern, E.; and Schubert, R. 2019. Beliefs in "brilliance" and belonging uncertainty in male and female stem students. *Frontiers in psychology*.

Essig, L. W. 2018. *A Content-Analytic Meta-Analysis of Gender Stereotyping in Screen Media*. Brigham Young University.

Feng, Y., and Shah, C. 2022. Has CEO gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36.

Ford, L. 2019. Geena Davis: damaging stereotypes on screen limit women's aspirations. https://www.theguardian.com/global-development/2019/oct/01/geena-davis-damaging-stereotypes-on-screen-limit-womens-aspirations. (accessed: 02.27.2023).

Garcia, C. 2016. Harold Cohen and AARON — a 40-year collaboration. https://computerhistory.org/blog/harold-cohen-and-aaron-a-40-year-collaboration/. (accessed: 02.27.2023).

Heikkilä, M. 2022. The viral AI avatar app Lensa undressed me—without my consent. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/. (accessed: 02.27.2023).

Hibbing, A. N., and Rankin-Erickson, J. L. 2003. A picture is worth a thousand words: Using visual images to improve comprehension for middle school struggling readers. *The reading teacher* 56(8).

Institute, G. D. 2016. Female characters in film and TV motivate women to be more ambitious, more successful, and have even given them the courage to break out of abusive relationships. https://seejane.org/gender-in-media-news-release/female-characters-film-tv-motivate-women-ambitious-successful-even-given-courage-break-abusive-relationships-release/. (accessed: 02.27.2023).

Invaluable. 2021. Art history timeline: Western art movements and their impact. https://www.invaluable.com/blog/art-history-timeline/. (accessed: 02.27.2023).

Javed, A. 2015. The media, the women and stem fields.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 3819–3828.

Kiefer, B. 2022. Nestlé brand is latest to venture into brave new world of AI art direction. https://www.adweek.com/creativity/nestle-brand-is-latest-to-venture-into-brave-new-world-of-ai-art-direction/. (accessed: 02.27.2023).

Koidan, K. 2020. 8 AI companies generating creative advertising content. https://www.topbots.com/ai-companies-generating-creative-advertising-content/. (accessed: 02.27.2023).

Lam, O.; Broderick, B.; Wojcik, S.; and Hughes, A. 2018. Gender and jobs in online image searches.

LeanIn.Org. 2023. You can't be what you can't see: The Lean In collection on Getty images. leanin.org/getty. (accessed: 02.27.2023).

Leslie, S.-J.; Cimpian, A.; Meyer, M.; and Freeland, E. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 347(6219):262–265.

Mattei, S. E.-D. 2022. Careful — Lensa is using your photos to train their AI. https://www.artnews.com/art-news/news/does-lensa-ai-use-your-face-data-for-selfies-1234649204/. (accessed: 02.27.2023).

Mello-Klein, C. 2022. The AI portrait app Lensa has gone viral, but it might be more problematic than you think. https://news.northeastern.edu/2022/12/09/portrait-ai-app/. (accessed: 02.27.2023).

Microsoft. 2023. Microsoft and OpenAI extend partnership. https://blogs.microsoft.com/blog/2023/01/23. (accessed: 02.27.2023).

Miller, C. C. 2014. LeanIn.org and Getty aim to change women's portrayal in Stock photos. https://www.nytimes.com/2014/02/10/business/leaninorg-and-getty-aim-to-change-womens-portrayal-in-stock-photos.html?smid=pl-share. (accessed: 02.27.2023).

Murnen, S. K.; Greenfield, C.; Younger, A.; and Boyd, H. 2016. Boys act and girls appear: A content analysis of gender stereotypes associated with characters in children's popular culture. *Sex roles*.

OpenAI. 2022. Reducing bias and improving safety in DALL·E 2. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2. (accessed: 02.27.2023).

Potter, W. J. 1993. Cultivation theory and research: A conceptual critique. *Human communication research* 19(4):564–601.

Q.ai. 2022. Dall·E Mini and the future of artificial intelligence art. https://www.forbes.com/sites/qai/2022/10/21/dalle-mini-and-the-future-of-artificial-intelligence-art/?sh=50e8121b7d78. (accessed: 02.27.2023).

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Rayner, A. 2016. Can Google's Deep Dream become an art machine? https://www.theguardian.com/artanddesign. (accessed: 02.27.2023).

Salkowitz, R. 2022. Midjourney founder David Holz on the impact of AI on art, imagination and the creative economy. https://www.forbes.com/sites/robsalkowitz. (accessed: 02.27.2023).

Schwemmer, C.; Knight, C.; Bello-Pardo, E. D.; Oklobdzija, S.; Schoonvelde, M.; and Lockhart, J. W. 2020. Diagnosing gender bias in image recognition systems. *Socius* 6:2378023120967171.

Shihadeh, J.; Ackerman, M.; Troske, A.; Lawson, N.; and Gonzalez, E. 2022. Brilliance bias in GPT-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)*.

Shrum, L. J. 1995. Assessing the social influence of television: A social cognition perspective on cultivation effects. *Communication Research* 22(4):402–429.

Shubber, K. 2013. Artificial artists: when computers become creative. https://www.wired.co.uk/article/can-computers-be-creative. (accessed: 02.27.2023).

Shutterstock. 2023. Shutterstock introduces generative AI to its all-in-one creative platform. https://www.prnewswire.com/news-releases/shutterstock-introduces-generative-ai-to-its-all-in-one-creative-platform-301729904.html. (accessed: 02.27.2023).

Srinivasan, R., and Uchino, K. 2021a. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Srinivasan, R., and Uchino, K. 2021b. Quantifying confounding bias in generative art: A case study. *arXiv preprint arXiv:2102.11957*.

Steele, C. 2022. Lensa AI is carrying gender bias into the future. https://www.pcmag.com/opinions/lensa-ai-is-carrying-gender-bias-into-the-future. (accessed: 02.27.2023).

Storage, D.; Charlesworth, T. E.; Banaji, M. R.; and Cimpian, A. 2020. Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology*.

Suddath, C. 2014. How Sheryl Sandberg and Getty are making Stock photos less sexist. https://www.bloomberg.com/news.

Sun, J.; Liao, Q. V.; Muller, M.; Agarwal, M.; Houde, S.; Talamadupula, K.; and Weisz, J. D. 2022. Investigating explainability of generative AI for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces*.

Sundararajan, L. 2021. Harold Cohen and AARON: Collaborations in the last six years (2010–2016) of a creative life. *Leonardo*.

Sung, M. 2022. The AI Renaissance portrait generator isn't great at painting people of color. https://www.artnews.com/art-news/news/does-lensa-ai-use-your-face-data-for-selfies-1234649204/. (accessed: 02.27.2023).

Ward, L. M., and Grower, P. 2020. Media and the development of gender role stereotypes. *Annual Review of Developmental Psychology* 2:177–199.

Welsh, J. 2013. These are the 7 things keeping women out of science careers. https://www.businessinsider.com/7-things-keeping-women-out-of-science-2013-10. (accessed: 02.27.2023).

Wiggers, K. 2021. Audit finds gender and age bias in OpenAI's CLIP model. https://venturebeat.com/business/audit-finds-gender-and-age-bias-in-openais-clip-model/. (accessed: 02.27.2023).

Crea.visions: A Platform for Casual Co-Creation with a Purpose Envisioning the Future through Human-AI Collaboration with Multiple Stakeholders

Janet Rafner
janetrafner@mgmt.au.dk
Center for Hybrid Intelligence
Aarhus University
Aarhus, Denmark

Blanka Zana
blanka.zana@gmail.com
Center for Hybrid Intelligence
Aarhus University
Aarhus, Denmark

Tristan Beolet
tristan.beolet@student-cs.fr
Center for Hybrid Intelligence
Aarhus University
Aarhus, Denmark

Safinaz Büyükgüzel
safinaz@mgmt.au.dk
Center for Hybrid Intelligence
Aarhus University
Aarhus, Denmark

Ewen Michel
ewen.michel@student-cs.fr
Center for Hybrid Intelligence
Aarhus University
Aarhus, Denmark

Neil Maiden
Neil.Maiden.1@city.ac.uk
City, University of London
Digital Creativity at London, England

Sebastian Risi
seb@itu.dk
Robotics, Evolution and Art Lab (REAL)
IT University of Copenhagen
Copenhagen, Denmark

Jacob Sherson
sherson@mgmt.au.dk
Center for Hybrid Intelligence
Aarhus University
Aarhus, Denmark

Abstract

With recent advances in Artificial Intelligence and increasing emphasis on human augmentation and collaboration, time is ripe for AI-enhanced support tools which empower the public to formulate and visualize a collective vision of societal issues such as climate change. Here, we report on crea.visions, a platform for human-AI co-creation within Sustainable Development Goals centered community engagement. We present in-the-wild experiments with four versions of crea.visions involving 1,000+ participants and 25,000+ generated images over three years: Versions 1 and 2 focused on developing the novel tool empowering citizens to artistically communicate their favorite abstract societal issues. In versions 3 and 4, the generic image generation GAN was replaced by custom-trained versions for Venice and Paris respectively. Refining the platform towards community-specific action, users of version 4 can geotag their identified problems, submit solutions candidates, and are actively linked up with relevant NGOs. Finally, version 4 includes the first workflow todate which combines AI image-generating modalities of sliders and text-to-image.

Introduction

The OECD (Schleicher 2010), the World Economic Forum (Belsky 2020) and many other international agencies argue that creativity is one of the top five skills in the 21st century. This includes how creativity is necessary for formulating and implementing a wide range of local and global solutions to complex issues such as the Sustainable Development Goals.[1] Thus, methods for fostering, improving, and facilitating human creativity have been studied for decades ranging from ideation interventions (Baas, De Dreu, and Nijstad 2008; Santanen et al. 2004) to creativity support tools (Frich et al. 2019; Sielis, Tzanavari, and Papadopoulos 2009). However, with the ever-improving Artificial Intelligence (AI) technologies, how AI can augment human creativity is becoming a prominent area for research and development. Human-AI co-creative systems, involve at least one human agent and one artificially intelligent agent collaborating with each other to build creative artifacts (Davis 2013; Kantosalo and Toivonen 2016). This collaborative activity has been defined as mutually influential contributions (Davis 2013), mixing human and computational initiatives (Yannakakis, Liapis, and Alexopoulos 2014) and the sharing of creative responsibility (Kantosalo and Toivonen 2016).

A recent trend called Casual Creators can be considered a subcategory of human-AI co-creative systems. Casual Creators promote efficient and enjoyable exploration of a possibility space, leading to the creation of unexpected artifacts that inspire feelings of pride, ownership, and creativity in the users who create them (Compton and Mateas 2015). At the core of these products often lie AI generators which empower amateur creativity by mapping simple low-dimensional input domains, e.g., sliders, to a complex high-dimensional output domain, e.g., images (Gajdacz et al. 2021). Beyond traditional entertainment purposes, researchers have demonstrated the utility of casual creators for creativity assessment (Rafner et al. 2020; Rafner 2021; Gajdacz et al. 2021), raising awareness of societal issues (Chang and Ackerman 2020; Luccioni et al. 2021; Rafner et al. 2021) and visualizing the future (Epstein, Schroeder,

[1] https://sdgs.un.org/goals

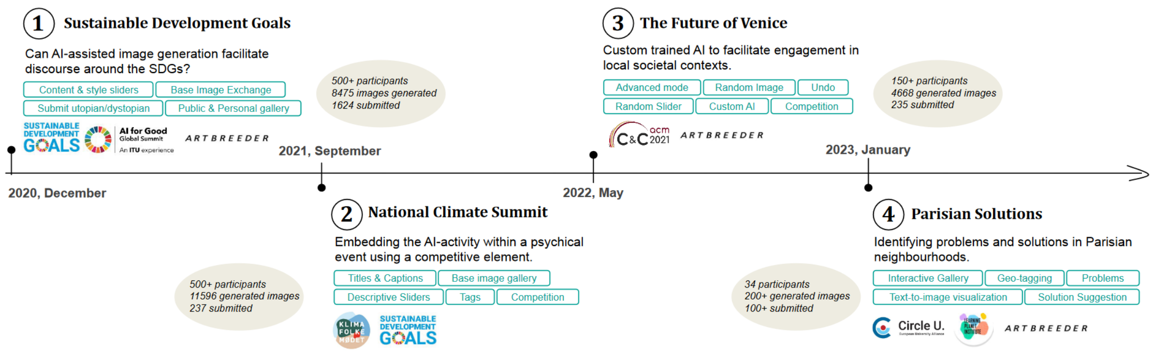


Figure 1: This figure presents a timeline of versions of crea.visions, the context in which it was used, the new key elements, stake holders, as well as number of participants and images generated and submitted for each of the four events.

and Newman 2022). These tools predominantly use Generative Adversarial Networks (GANs), a well-known machine-learning model designed to produce artificial images that are nearly indistinguishable from real images (Borji 2019). GAN's abilities present two major advantages when it comes to designing co-creative systems: the generation is deterministic, i.e., non-random, and the generative part of the network is continuous and locally coherent over the latent space, [2] both properties making it intuitive and controllable to explore for naive users.

Despite these promising developments, there remains a significant research gap. Current literature on casual creators used for research purposes is still in its infancy; the articles cited above primarily focuses on prototype descriptions or pilot studies of user interactions, with little attention given to large-scale multi-stakeholder studies (e.g., participants, government officials, companies) or successive iterations of a tool. In other words, the potential of Casual Creators to contribute to large-scale, public use for various research and societal applications remains largely unexplored and untested in real-world settings.

To address this gap, we propose an approach grounded in the Community Citizen Science (CCS) framework (Hsu and Nourbakhsh 2020). Our objective is to move beyond prototype and pilot studies, and extend the application of Casual Creators into large-scale, ecological environments with diverse stakeholders. The CCS framework, emphasizing participatory democracy and community co-design, offers a promising route towards achieving this goal. In the context of Human-Computer Interaction, CCS has been utilized to enhance scientific research and community empowerment (Hsu and Nourbakhsh 2020). Examples include mobile apps that allow residents to track pollution odors or generate high-resolution landscape imagery (Hsu and Nourbakhsh 2020). However, the application of CCS principles to human-AI co-creative systems remains largely unexplored. This study aims to address this uncharted territory, presenting the first known application of CCS principles in this context. We believe this approach can yield significant benefits, opening up new possibilities for the use and advancement of Casual Creators.

[2] n-dimensional vector space from which the noise vectors used for image generation are sampled. Notions of latent space and decoder/generator tend to be used in similar contexts.

This article presents the development of four versions of crea.visions, a platform for human-AI co-creation designed for community involvement and social good; see Fig. 1 for an overview on the four versions. We aim to show how combining human-AI co-creative technology and CCS principles holds potential for new forms of public engagement on complex socio-environmental topics such as climate change. We approached the development of crea.visions with design goals (DG) of a user-friendly and intuitive interface (DG1), meaningful content and motivation for activities (DG2), meeting technical requirements (DG3), and community buy-in and co-design with multiple stakeholders (DG4). We provide an overview of the platform, describe and analyze data from four experiments over three years. We conclude with a discussion on future directions for crea.visions.

Crea.visions System Overview

The core mechanics of crea.visions allows users to blend 'style' (small scale features and texture) and 'content' (large scale features) components of a set of source images into new images (see Fig. 2). Images are generated by the pre-trained StyleGAN2 (Karras et al., 2020) and the system is developed in the unity game engine.

Crea.visions enables a form of alternating, turn based, co-creativity where first the human provides an input and then the model generates an image based on the users input. The interface is simple with minimal features to allow for very brief on-boarding instructions before users can generate contextually meaningful images.

Crea.visions Version 1: Sustainable Development Goals, Online Data Collection

The main purpose of V1 was to demonstrate the technical viability of crea.visions (DG3) and the problem framing needed to engage the public in creating thought-provoking visions of utopias and dystopias in order to raise awareness of socioscientific problems related to the SDGs. In this experiment, as with the others, IRB approval was received

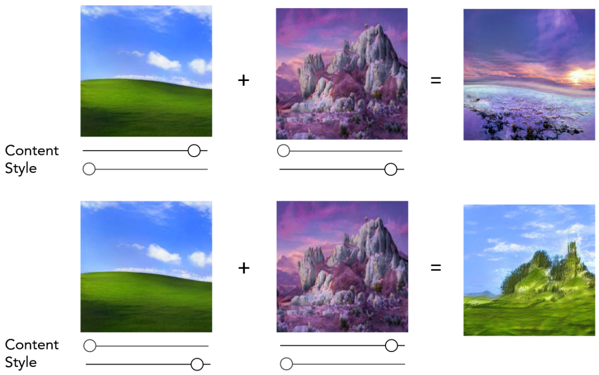


Figure 2: This image illustrates how in crea.visions, players use two sliders per image in order to control the 'style' (small scale features and texture) and 'content'(large scale features) they want each image to contribute into a new image.

from Aarhus University. Participants gave informed consent prior to participation.

Multi-Stakeholder Alignment and Experimental Design

V1 was developed with Artbreeder, a popular online image generation platform, and the United Nations online platform, AI4good. The authors met regularly over the course of six months with the representatives from both platforms to optimize usability. Due to Covid19 the launch was online only. The use of crea.visions was promoted by all stakeholders via social media channels including Twitter, Facebook, LinkedIn, and Instagram. Participation was completely voluntary. Data was collected between December 2020 and March 2021 and included slider movements and user clicks, and submitted images.

Technical Considerations

The preexisting StyleGAN model (Karras et al. 2020) was used in V1. The roughly 200 base images originally from the Artbreeder community were selected by the first author with the criteria of diversity in style, content, color, and motif to facilitate a wide range of possible utopian and dystopian images. The V1 experiment also served to test the capacity of the tool (i.e., number of simultaneous users) as image blending requires heavy computational power.

Gameplay and User Experience

Participants were presented with five options on the landing page: a tutorial, image blending, a public gallery, next steps (project information), and a personal gallery. Participants were given the challenge to imagine the world 50 years from now and generate images of possible futures. Given the content of the images, they were particularly effective for climate related issues (fires, flooding, etc.). The user flow guided first-time users to begin with the tutorial, explore the public gallery for inspiration, then proceed to blend up to four images which they could change by clicking on them. A click on a base image randomly loaded a new one. When a player finished creating their image, they could submit the blended image, tagging it as either utopian or dystopian. The submitted images were anonymously published in a public gallery for voting (Fig 3, Right) and could be downloaded or further edited by other participants. The personal gallery stored all images created during the individual session.

Descriptive Statistics, Exploratory Analysis, and Lessons Learned

In total there were 580 user sessions and 8475 images generated. Exploratory analysis from this study is published in (Rafner et al. 2021), showing that V1 allowed users to create images that express both anxiety and hope for the future, affirmed that user-generated images express these ideas in ways that are meaningful to others, and began to investigate which specific features of images (color, motif, style) are more closely related to dystopian or utopian ideas. The research explored in (Rafner et al. 2021) presents results on image analysis of the generated images from V1 where as the present article focuses much more on the system design user experience, and multi-stake holder alignment for V1-V4. The testing of V1 also identified that one virtual machine with an Nvidia A100 GPU, 4 cores and 25GB RAM could support approximately 50 simultaneously users. The partners were happy with the outcome of the launch as a proof of concept, however as suggested by the CCS Framework, future iterations needed to focus on co-creating features with multiple stakeholders (DG4), contextualizing the activities to bring more meaning and adding more options for the participants for expressing intentions (DG2).

Crea.visions Version 2: National Climate Summit: In-Person Events with Students and the General Public

The main purpose of V2 was to improve user control, specifically participants' ability to describe their intentions (DG1) and to evaluate if a competition component was beneficial for crea.visions player motivation (DG2).

Multi-stakeholder Alignment and Experimental Design

In September 2021, V2 was launched at a Danish People's Climate Summit[3] in Middelfart, Denmark; two data collection sessions were conducted. Participants were instructed to create their vision of the future 50 years from now using V2. Log data, titles, captions, and tags were collected and participants provided qualitative feedback. Participants were incentivized to relate their work to environmental topics, in particularly climate change, while also specifically "having to picture their view of the future." The first data collection was with 500 high school students who were participated during in-class time, working in teams to produce over 170 submissions over one week; students used their own laptops. The authors met with coordinators from the

[3]Klimafolkemødet at https://klimafolkemoedet.dk/

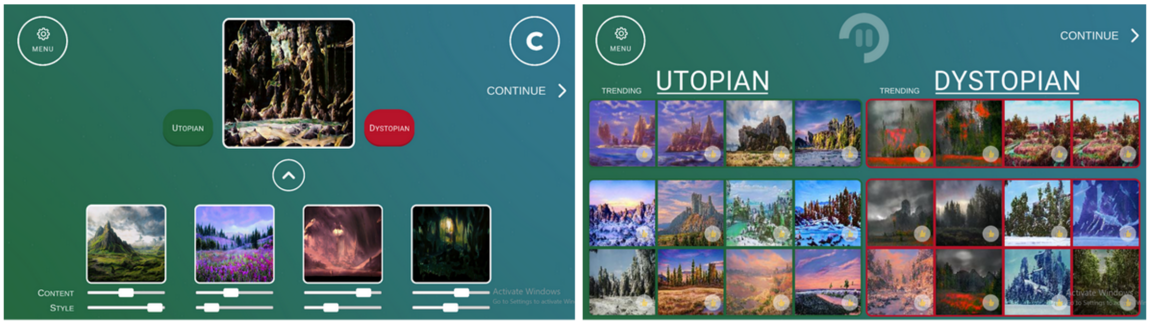


Figure 3: V1. Left: Blending interface. Right: Public gallery of utopian and dystopian images.

Figure 4: V2: Photos from climate summit event.

school multiple times prior to the event to prepare an introductory presentation on human-AI co-creativity for the students. The second data collection took place at the public climate summit event, which hosted numerous environment-related speakers and activities. Participants interacted with V2 on laptops, available at open stands (Fig 4). Participation in the event was advertised via social media by all partners and participation was on a voluntary basis. However, participants could submit their creations to a competition with a prize of 1350 EUR. Submissions were evaluated by an expert panel based on the novelty and appropriateness. Titles and captions played a crucial role in the competition, as they allowed the intentions of the player to be described (Fig 5, Right).

At the climate summit's closing ceremony, the mayor of Middlefart presented the award. The winning submission was made by a group of three students, who commented "It was fun to try something new where you could express yourself creatively." A highschool teacher part of the organizing team said that crea.visions was "very user-friendly, and quite easy to navigate" and that "for me, and for my colleagues, I think the most positive thing has been that it is so different from what we are used to in everyday life". The leader of the climate summit concluded that "we had a really good collaboration with the team of crea.visions, and I expect that we will continue with next year's meeting".

Technical Considerations

The GAN used and image selection was identical to V1.

Gameplay and User Experience

Participants were first shown a video which explained the game and competition. They then proceed to blend images. V2 was equipped with a two-step submission page that allowed participants to provide a title, a description and tags (Fig 5, Left), and rate their image properties on various scales, (e.g. natural/man-made, permanent/reversible - see Fig. 6). The new submission process was designed to prompt deeper reflection in the players regarding various dimensions of the image, and add narrative to the submitted images. The base image swap feature was developed into a controlled selection process, where upon a click on a base image a gallery of images would open, allowing people to browse through various options (Fig. 6) as opposed to the random cycling through of images, supported by V1. Similarly to V1, V2 had a social library function and personal user gallery.

Descriptive Statistics, Exploratory Analysis, and Lessons Learned

In total, 11 596 images were generated and 237 were submitted to the competition. Exploratory data analysis was performed to investigate how participants explored the possibility space before submitting an image (e.g., are they thoughtful creations or simply random submissions) to help determine if V2 was appropriately scaffolding the interaction (DG1). Users generated on average 43 images before submitting, with a range of 0 to 272. The high number of generated images before submission was interpreted as a proxy for engagement: if the submission was preceded by a high number of iterations, it could be considered a thoughtful creation. On average, the highest ranked image creators (N=16) explored more than other players, however as the sample size of the winners group was small, no significant conclusions can be drawn regarding group differences. Qualitative feedback indicated it was difficult for the players to know how much of the image space they had explored and were frustrated that they were unable to go back to previously generated images. By performing sentiment analysis on titles and captions, we found that people's dystopian/utopian

Figure 5: V2. Left: image blending interface Middle: The winner receiving the prize. Right: The winning submission.

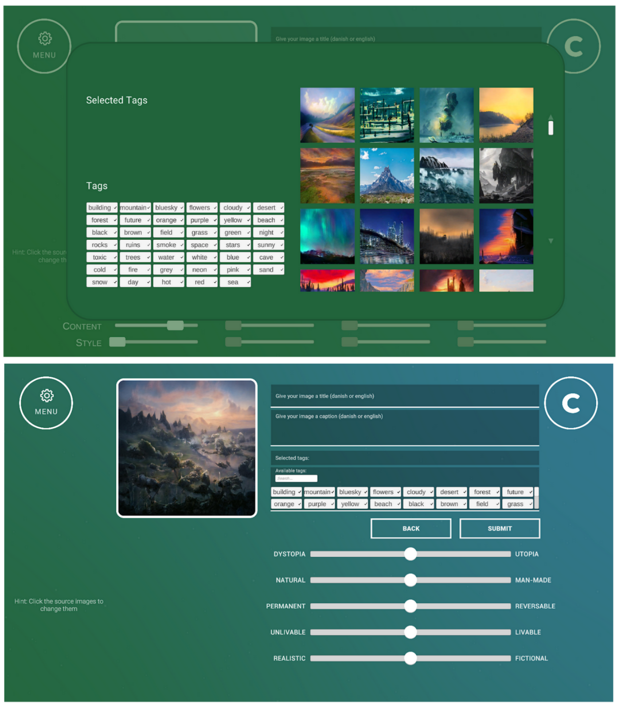


Figure 6: Screenshots from V2. Top: Base image Gallery. Bottom: Second Submission Page.

slider usage was positively correlated (r=.24, p=.009) with the sentiment of their captions (e.g., the more utopian an image was rated on the scale, the more positive the caption's sentiment was), indicating player intentions are reflected through both slider use and captions. It was observed that a thoughtful submission took at least 20mn. The gallery feature and the 'upvoting' was intended to provide a shorter, yet meaningful interaction for those with less time, but qualitative feedback indicated this was not the case (DG3).

Logistically speaking we also observed that in order to make a thought-through submission, participants took at least twenty minutes and required a comfortable place to sit and concentrate, out of the sun to avoid glare on their screens.

Crea.visions Version 3: The Future of Venice: Venetian Residents and Tourists

The main purpose of V3 was to use custom images for local specificity (DG2, DG4) and usability improvements to allow the participants to navigate through the GAN space more fluently (DG1, DG3).

Multi-Stakeholder Alignment and Experimental Design

V3 was developed for the 2022 Creativity & Cognition Conference in Venice, Italy. As part of the crea.visions week event, pop-up workshops were held on the streets of Venice. Log data from the blending, the titles, and captions was saved. Participants were asked to provide general qualitative feedback about their experience. Participants' ages ranged from 16-70+. Through social media, word of mouth, and on-the-street recruitment, the public was invited to create images of the future of the city in order to spark dialogues and self-reflection around sustainability, climate change, and tourism. As recommended by the CCS framework we coordinated with our local collaborators, including the General Co-Chair of the conference; we visited locations that attract a variety of groups, such as cafes, schools, parks, and local communities. Participation was voluntary, but participants were provided with refreshments while they sat and generated the images. Participants could also choose to submit their visions (images with descriptions) to a competition; winning visions were exhibited from 22-25 June 2023 as part of the conference art exhibition. Visions were evaluated in the same manner as V2. At the exhibition, we provided printed-out take-away postcards of 16 selected images.

Figure 7: V3: Pictures from the Venice event.

Figure 8: Left:V3 blending interface, advanced mode toggled on. Middle: Photograph from crea.visions Week. Right: selected images in postcard format including the final image, source images, title, caption and creators name.

Technical Considerations

Unlike V1 and V2 which used a generic pre-trained GAN, our Venetian partners wanted a version that produced recognizable Venetian architectural features to be contextually meaningful to the target group (DG2, DG4). Thus, the StyleGAN model was retrained (Varkarakis, Bazrafkan, and Corcoran 2020) on images of Artbreeder and Venetian buildings (see Fig. 9). In order to maintain the diversity of the pre-trained StyleGAN model, and achieve the continuous building-like latent space needed for this application, the standard training procedure for StyleGAN was used, even though it is computing intensive and required high-end hardware. The model requires 12 GB of VRAM, and common consumer-grade graphic cards often have between 4 and 8 GB of VRAM so a fifth of a Nvidia A100 has been used for this purpose. The retraining set was composed of 5026 images, 1855 of which were crowdsourced photos taken of Venice by our local partners. The rest of the images were generated by the original model. To bias the model to produce more images with certain features (vegetation, water, etc.) 16 base images containing the desired features were selected by the first author. These were blended together in order to achieve a reasonably sized dataset with sufficient variations:

Figure 9: Image illustrating the GAN's blending process for training image generation in V3. The far left and right images are the source images, and the images in the middle represent images that could be blended using these two images with various slider settings.

Gameplay and User Experience

Participants received a verbal introduction to the tool and then blended images using laptops pre-loaded with V3. In V3, the blended image is in the middle of the interface, and the source images are in the corners (Fig. 8, Left) to accommodate the advanced mode (described below). After blending images, participants moved to the submission page, giving their image a title, caption, and tags. The base image selection gallery was carried on from V2. New features were added, such as a random images button, which set four random base images and a random sliders button changed all four images' slider settings in order to facilitate the participants' exploration process (DG1). An undo button was added, allowing the user to return to the previous image settings as well as a button enabling the advanced mode (See Fig. 8, Left), displaying 16 images around the blended image, each the result of slightly adjusted slider settings compared to the current image. These served as a preview of possible images. Once a preview image was clicked on, the respective sliders would change and the selected image would be generated.

Descriptive Statistics, Exploratory Analysis, and Lessons Learned

4668 images were generated and 235 submitted. We were pleasantly surprised by the positive reception of V3 by Venetians aged 60+, many of whom voiced that they used little technology. A Venetian marketing student said, "It can give the people a great impression of the future. Not just an image in your mind, but something you can see on the laptop. So you can create it and actually see it." Multiple participants mentioned that the platform was "easy and intuitive" however, image content variety (more people, vegetation) and aspects of usability (text size, tool compatibility) could be improved. During the sessions, it became apparent that the two-step submission process did not lend itself intuitively to the generation of multiple submissions. The introduction of the second screen made the submission process feel rather final and gave people the impression that they finished a task. Additionally, due to the additional computational load of generating the 16 supporting images in every step, using the specifications from V1 and V2 could support approximately 3 simultaneous users, making scalability of the function difficult.

Crea.visions Version 4: Parisian Solutions: Participatory Design, Text-to-Image and an Online Universe

The main purpose of V4 is to provide support for civic problem solving (DG2, DG4), add meaningful short interactions

(DG1,DG2), and to include text-to-image technology to expand creation possibilities (DG1).

Multi-stakeholder Alignment and Experimental Design

V4 was developed for the 2023 Learning Planet Festival, held in January in Paris, France. The events included three in person and two virtual workshops with bachelor and masters students at the Learning Planet Institute (LPI) and the general public. The events were advertised on both the Learning Planet Festival as well as Learning Planet Institute websites and social media. The new features and user flow was determined after holding two participatory design (Sanders, Brandt, and Binder 2010) workshops (N=19) in November 2022 with Parisians. Participatory design is a method underscored by the CCS framework to engage community stakeholders in designing the tools and interventions as opposed to simply participating in or with a final version (DG4). Participation in all events was voluntary, but during image creation participants were provided with refreshments.

Technical Considerations

Similar to V3, the GAN has been retrained based on crowdsourced photographs from Paris and stock images to ensure diversity and representation of landmarks in Paris. V4 introduces an online companion app extending the features of the game (described below). To alleviate the load on the GPU node, in the companion app the blender is replaced with an interactive gallery of pre-generated Paris-GAN images with progressive navigation towards preferred features (Fig. 10). Additionally, due to the new open source availability of text-to-image generation, ***we designed the first workflow todate which combined image blending modalities of both sliders and text-to-image*** with PlaygroundAI (https://playgroundai.com/)

Gameplay and User Experience

Participants were introduced to the centeral concepts in human-AI co-creativity, V4 of crea.visions and then prompted to define a problem and identify where in Paris it is severe. Then they generated an image of the problem using sliders, then refined their image by feeding it as input to a text-to-image stable diffusion model (Borji 2022) using PlaygroundAI. Participants then gave their image a title, and brainstormed solutions to the problem and identified which existing NGOs could get people engaged to help solve the problem. Participants could work individually or in pairs. Based on additional user testing from V3, the advanced mode was classified as confusing, adding little value to the user experience, thus we decided to remove it to improve the game flow (DG1). To extend its reach V4 introduces an online companion app allowing users to create similar submissions more rapidly than in the game (DG1). Users could locate their problem and/or solution on an interactive map and rank and discuss each other's submissions. The companion app lists online submissions alongside ingame submissions in a submission gallery that can be filtered by problem category.

Descriptive Statistics, Exploratory Analysis, and Lessons Learned

There were approximately 100 participants in the workshops and 34 submissions. See Fig. 12 for an example submission. Both images and log data from the blending process in crea.visions were collected, as well as all images generated in the text-to-image tool including prompts. As a whole, participants enjoyed the interactive aspect of the study and found the text-to-image tool in addition to the slider based crea.visions helpful in exploring and refining ideas. A survey was administered to the participants, which included questions about their prior experience with image blending tools and their overall experience with using the tools in a creative problem-solving setting. The results showed an equal distribution of answers between those who had prior experience with image blending tools and those who did not. The most popular topic that participants addressed with the tools was the environment. 50% of the participants reported being very interested in the problem they selected. Based on the survey, the tools were found to be helpful in visualizing and supporting idea exploration, as well as improving the quality of the results. For example, one participant commented, "Crea.visions made it easier to see how things would look like, if the problem was tackled properly." Weaknesses of the workshops included the lack of integration between the tools (e.g., participants had to move between crea.visions, Playground AI, and the companion app), the lengthy questionnaire, as well as the difficulty in prompt engineering.

Discussion and Future Work

The main novelty of this platform is its combination of human-AI co-creative technology with principles from the CCS framework to engage the public in image generation enabling civic expression and communication of societal challenges such as climate change. We approached the development of crea.visions with design goals of a user-friendly and intuitive interface (DG1), meaningful content and motivation for activities (DG2), meeting technical requirements (DG3), and community buy-in and co-design with multiple stakeholders (DG4). With respect to the usability improvements for exploring the vast GAN space (DG1, DG3), initial user exploration was supported through a single button generating new random images rather than manually adjusting each slider and base image. Furthermore, an experiment was conducted with dynamic visualization of the immediate possibility space around the current solutions by presenting 16 close-lying images. While this feature was useful to many participants, it was also computationally costly and distracting to others. As such, it may be reintroduced in future versions but was abandoned for now. Instead, the powerful text-to-image technology was explored for the V4 (DG1, DG3).

In terms of enriching the submission format, it was found that images could not stand alone, so they were supplemented with titles and short narratives (added in V2), which were often compelling, emotional and artistic. In V4, the submission format was enriched with geotagging, solution

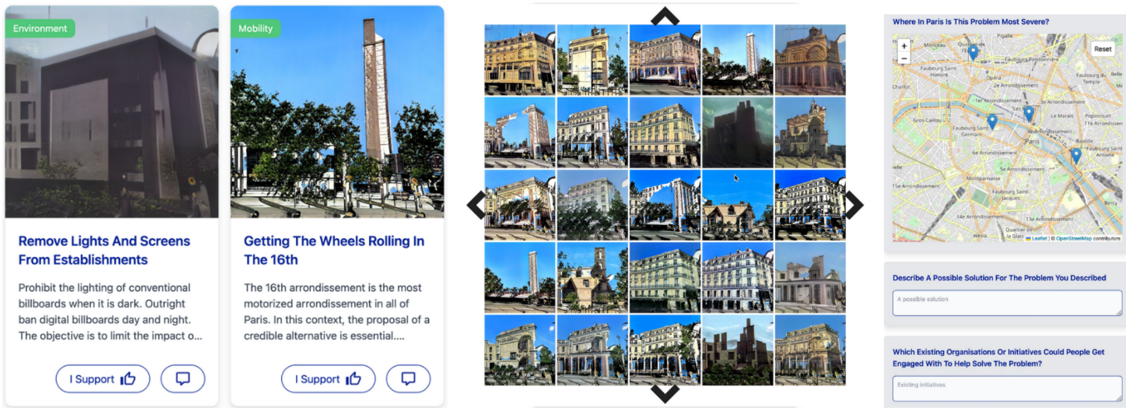


Figure 10: Screenshots from V4 companion app. Left: Submission Cards with tags from the submissions gallery. Middle: Interactive GAN image selector. Right: Interactive map and prompt fields in the submission form.

Figure 11: Example image created during the Paris workshops. Participant's Caption: The concept is a new way to move in a city. We don't use the roads because there are a lot of vehicles. That's a ecological transport and we choose the most used roads to take the most people possible..

definition, NGO contact details and a chat feature. Exploratory quantitative and qualitative analysis of the submissions in each version have been performed, providing initial insights into the type of analysis that could be done in the future.

In coming iterations, the trend towards richer and smoother interactions as well as increased and lasting community impacts will be continued (DG2, DG4). This includes travelling exhibitions and using the platform as personal and approachable initiators of complex value-based discussions. A concrete proposed activity is having policy makers start their panel debates with generated images and their personal visions of the future. Concretely, in 2023, we are planning to run events in the Botanical Gardens in Belgrade, Serbia and at the Central Library, in Aarhus Denmark.

The potential of casual creators with a purpose goes beyond their immediate application in the crea.visions project. They can serve as powerful conversation starters and debate initiators across various settings and demographics, from children and the elderly to individuals from different languages and cultures. By integrating AI with human creativity and applying the CCS principles, our work indicates that one can facilitate inclusive and insightful discussions about societal and climate related topics which could also be useful in numerous fields, from education to policy-making.

Conclusion

In conclusion, this article presented the development and evolution of crea.visions for human-AI co-creation, specifically for community engagement and social good. Through four independent experiments, the platform was tested with over 1,000 participants and 25,000 generated images. The platform has evolved over the course of three years to become increasingly tailored for community-specific action, with the most recent version allowing users to geotag problems within a specific part of a city and submit solutions that are linked up with existing NGOs. The article emphasized that this platform is a step towards empowering the public to contribute to building and visualizing a collective vision of current societal issues and possibilities for our future world through the use of AI enhanced support tools. We hope that this platform inspires research institutions, local authorities and civil society organizations across the world to partner up to continue developing AI-tools for community empowerment, and strengthening the link between science and civil society.

Acknowledgements

The authors would like to thank SHAPE, Sonophilia Foundation, CircleU, Carlsberg Foundation and Novo Nordisk Foundation for their generous support of this work.

References

Baas, M.; De Dreu, C. K.; and Nijstad, B. A. 2008. A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological bulletin* 134(6):779.

Belsky, S. 2020. Creativity will be key to competing against ai in the future workforce–here's how. In *World Economic Forum. Retrieved*, volume 12, 10–20.

Borji, A. 2019. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding* 179:41–65.

Borji, A. 2022. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*.

Chang, J., and Ackerman, M. 2020. A climate change educational creator. In *ICCC*, 77–80.

Compton, K., and Mateas, M. 2015. Casual creators. In *ICCC*, 228–235.

Davis, N. M. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Ninth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Epstein, Z.; Schroeder, H.; and Newman, D. 2022. When happy accidents spark creativity: Bringing collaborative speculation to life with generative ai. *arXiv preprint arXiv:2206.00533*.

Frich, J.; MacDonald Vermeulen, L.; Remy, C.; Biskjaer, M. M.; and Dalsgaard, P. 2019. Mapping the landscape of creativity support tools in hci. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–18.

Gajdacz, M.; Rafner, J.; Langsford, S.; Hjorth, A.; Bergenholtz, C.; Biskjaer, M. M.; Noy, L.; Risi, S.; and Sherson, J. F. 2021. Crea. blender: A gan based casual creator for creativity assessment. In *ICCC*, 405–409.

Hsu, Y.-C., and Nourbakhsh, I. 2020. When human-computer interaction meets community citizen science. *Communications of the ACM* 63(2):31–34.

Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the Seventh International Conference on Computational Creativity*, 77–84.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Luccioni, A.; Schmidt, V.; Vardanyan, V.; and Bengio, Y. 2021. Using artificial intelligence to visualize the impacts of climate change. *IEEE Computer Graphics and Applications* 41(1):8–14.

Rafner, J.; Hjorth, A.; Risi, S.; Philipsen, L.; Dumas, C.; Biskjær, M. M.; Noy, L.; Tylén, K.; Bergenholtz, C.; Lynch, J.; et al. 2020. Crea. blender: a neural network-based image generation game to assess creativity. In *Extended abstracts of the 2020 annual symposium on computer-human interaction in play*, 340–344.

Rafner, J.; Langsford, S.; Hjorth, A.; Gajdacz, M.; Philipsen, L.; Risi, S.; Simon, J.; and Sherson, J. 2021. Utopian or dystopian?: using a ml-assisted image generation game to empower the general public to envision the future. In *Creativity and Cognition*, 1–5.

Rafner, J. 2021. Creativity assessment games and crowdsourcing. In *Creativity and Cognition*, C & C '21. New York, NY, USA: Association for Computing Machinery.

Sanders, E. B.-N.; Brandt, E.; and Binder, T. 2010. A framework for organizing the tools and techniques of participatory design. In *Proceedings of the 11th Biennial Participatory Design Conference*, PDC '10, 195–198. New York, NY, USA: Association for Computing Machinery.

Santanen, E. L.; Briggs; O, R.; Vreede; and De, G.-J. 2004. Causal relationships in creative problem solving: Comparing facilitation interventions for ideation. *Journal of management information systems* 20(4):167–198.

Schleicher, A. 2010. The case for 21st-century learning - oecd. Accessed: 2021-04-14.

Sielis, G. A.; Tzanavari, A.; and Papadopoulos, G. A. 2009. Enhancing the creativity process by adding context awareness in creativity support tools. In *International Conference on Universal Access in Human-Computer Interaction*, 424–433. Springer.

Varkarakis, V.; Bazrafkan, S.; and Corcoran, P. 2020. Re-training stylegan-a first step towards building large, scalable synthetic facial datasets. In *2020 31st Irish Signals and Systems Conference (ISSC)*, 1–6. IEEE.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. *asdf*.

Towards Mode Balancing of Generative Models via Diversity Weights

Sebastian Berns ○, **Simon Colton** ○ and **Christian Guckelsberger** △○
○ School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
△ Department of Computer Science, Aalto University, Espoo, Finland

Abstract

Large data-driven image models are extensively used to support creative and artistic work. Under the currently predominant distribution-fitting paradigm, a dataset is treated as ground truth to be approximated as closely as possible. Yet, many creative applications demand a diverse range of output, and creators often strive to actively diverge from a given data distribution. We argue that an adjustment of modelling objectives, from pure mode coverage towards mode balancing, is necessary to accommodate the goal of higher output diversity. We present *diversity weights*, a training scheme that increases a model's output diversity by balancing the modes in the training dataset. First experiments in a controlled setting demonstrate the potential of our method. We discuss connections of our approach to diversity, equity, and inclusion in generative machine learning more generally, and CC specifically. An implementation of our algorithm is available at https://github.com/sebastianberns/diversity-weights

Introduction

Large image generation models (LIGMs), in particular as part of text-to-image generation systems (Ramesh et al., 2021; Saharia et al., 2022), have been widely adopted by visual artists to support their creative work in art production, ideation, and visualisation (Ko et al., 2023; Vimpari et al., 2023). While providing vast possibility spaces, LIGMs, trained on huge image datasets scraped from the internet, not only adopt but often exacerbate data biases, as observed in word embedding and captioning models (Bolukbasi et al., 2016; Zhao et al., 2017; Hendricks et al., 2018). The tendency to emphasise majority features and to primarily reproduce the predominant types of data examples can be limiting for many computational creativity (CC) applications that use machine learning-based generators (Loughran and O'Neill, 2017). Learned models are often used to illuminate a possibility space and to produce artefacts for further design iterations. Examples range from artistic creativity, like the production of video game assets (Liapis, Yannakakis, and Togelius, 2014; Volz et al., 2018), over constrained creativity, e.g. industrial design and architecture (Bradner, Iorio, and Davis, 2014), to scientific creativity, such as drug discovery (Madani et al., 2023). Many of these and similar applications would benefit from higher diversity in model output. Given that novelty, which underlies diversity, is considered one of the essential aspects of creativity (Boden, 2004; Runco and Jaeger, 2012), we expect that, vice versa, a stronger focus on diversity can also foster creativity (cf. Stanley and Lehman, 2015).

Most common modelling techniques, however, follow a distribution-fitting paradigm and do not accommodate the goal of higher diversity. Within this paradigm, one of the primary generative modelling objectives is *mode coverage* (Zhong et al., 2019), i.e. the capability of a model to generate all prominent types of examples present in a dataset. While such a model can in principle produce many types of artefacts, it does not do so reliably or evenly. A model's probability mass is assigned in accordance to the prevalence of a type of example or feature in a dataset. Common examples or features have higher likelihood under the model than rare ones. As a consequence, samples with minority features are not only less likely to be obtained by randomly sampling a model, they are also of lower fidelity, e.g. in terms of image quality. Related studies on Transformer-based language models (Razeghi et al., 2022; Kandpal, Wallace, and Raffel, 2022) have identified a "superlinear" relationship: while training examples with multiple duplicates are generated "dramatically more frequently", examples that only appear once in the dataset are rarely reproduced by the model.

In this work, we argue for an adjustment of modelling techniques from mode coverage to *mode balancing* to enrich CC with higher output diversity. Our approach allows to train models that cover all types of training examples and can generate them with even probability and fidelity. We present a two-step training scheme designed to reliably increase output diversity. Our technical contributions are:

- *Diversity weights*, a training scheme to increase a generative model's output diversity by taking into account the relative contribution of individual training examples to overall diversity.
- *Weighted Fréchet Inception Distance (wFID)*, an adaptation of the FID measure to estimate the distance between a model distribution and a target distribution modified by weights over individual training examples.
- A proof-of-concept study, demonstrating the capacity of our method to increase diversity, examining the trade-off between artefact typicality and diversity.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

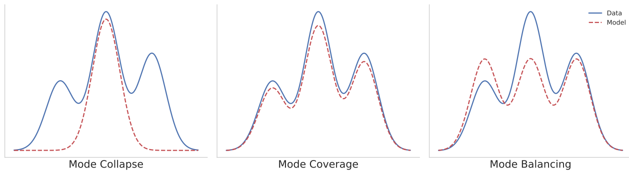


Figure 1: *Mode collapse*: the model does not cover all modes in the data distribution. *Mode coverage*: the data distribution's modes are modelled as closely as possible w.r.t. their likelihood. *Mode balancing*: the model covers all modes, but with equal likelihood.

In the following sections, we first introduce the objective of *mode balancing* and highlight its importance for CC based on existing frameworks and theories. Then, we provide background information on the techniques relevant for our work. Next, we present our *diversity weights* method in detail, as well as our formulation of *Weighted FID*. Following this, we present the setup and methodology of our study and evaluate its results. In the discussion section, we contribute to the debate on issues of diversity, equity, and inclusion (DEI) in generative machine learning more generally, and CC specifically, by explaining how our method could be beneficial in addressing data imbalance bias. This is followed by an overview of related work, our conclusions and an outlook on future work.

Mode Balancing

Generative deep learning models now form an integral part of CC systems (Berns et al., 2021). A lot of work on such models is concerned with *mode coverage*: to match a data distribution as closely as possible by accurately modelling all types of examples in a dataset (fig. 1). In the specific case of generative adversarial networks (GANs), great effort is put into preventing *mode collapse*, a training failure state in which a model disregards important modes and is only able to produce a few types of training examples. Mode coverage is captured formally in common evaluation measures such as Fréchet Inception Distance (FID) and Precision–Recall (PR). Crucially, this is always done in reference to the training set statistics or data manifold. In this context, diversity is often arguably misused to refer to mode coverage. While mode coverage describes the fraction of modes in a dataset that are represented by a model, the diversity of a model's output, if understood more generally and intuitively, can theoretically be higher than that of the dataset.

Mode coverage is conceptually similar to the notion of *typicality* (Ritchie, 2007). Defined as the extent to which a produced output is "an example of the artefact class in question", a model which only generates outputs with high typicality, if sampled at random, has to provide most support to those training set examples with the highest density of features characteristic of that artefact, i.e. to maximise mode coverage. Crucially, sampling from the model would resemble going along the most well-trodden paths in the possibility space defined by the dataset and, as Ritchie already suggests, counteract novelty as a core component of creativity (Boden, 2004; Runco and Jaeger, 2012).

Crucially, mode balancing breaks with the convention of viewing the dataset as 'ground truth'. Instead, we consider the dataset to provide useful domain information and the characteristics of *typical* examples (Ritchie, 2007). But a data distribution does not have to be matched exactly. Particularly in artistic applications, creators often strive to *actively diverge* from the typical examples in a dataset (Berns and Colton, 2020; Broad et al., 2021). To stay with our metaphor, borrowed from Veale, Cardoso, and Pérez y Pérez (2019), *mode balancing* allows us to walk more along the less trodden paths and thus especially support exploratory and transformational creativity (Boden, 2004; Stanley and Lehman, 2015). In contrast to the mode coverage paradigm, in mode balancing, diversity is measured independently of the training data distribution. In the theoretical case of a balanced dataset of absolutely dissimilar examples, i.e. multiple equally likely modes, our method would assign uniform weights to all examples and thus be identical to standard training schemes with random sampling.

Background

Probability-Weighted Vendi Score

We adopt the Vendi Score (VS) as a measure of dataset diversity and employ its probability-weighted formulation in our work (Friedman and Dieng, 2022). Given a set of artefacts $x_1, \ldots, x_n$, the probability-weighted VS is based on a probability vector $\mathbf{p} = (p_1, \ldots, p_n)$ and a similarity matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ between pairs of artefacts such that $\mathbf{K}_{ii} = 1$. Calculating VS involves various steps. First, the probability-weighted similarity matrix is defined as $\mathbf{K}^{\mathbf{p}} = \operatorname{diag}(\sqrt{\mathbf{p}})\,\mathbf{K}\operatorname{diag}(\sqrt{\mathbf{p}})$. Its eigenvectors $\lambda_1, \ldots, \lambda_n$ can be obtained via the eigendecomposition $\mathbf{K}^{\mathbf{p}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$, where $\boldsymbol{\lambda} = \operatorname{diag}(\mathbf{\Lambda})$. The probability-weighted Vendi Score (VS) is the exponential of the Shannon entropy of the eigenvalues of the probability-weighted similarity matrix:

$$\mathrm{VS}(\mathbf{K}, \mathbf{p}) = \exp\Big(- \sum^{n} \lambda_i \, \log \lambda_i \Big) \tag{1}$$

Also known as *perplexity*, exponential entropy can be used to measure how well a probability model predicts a sample. Low perplexity indicates good prediction performance. Consequently, the more diverse a sample, the more difficult its prediction, the higher the perplexity and its VS.

Illustrative Example The probability vector p represents the relative abundances of individual artefacts. Instead of repeating identical artefacts in a set, their prevalence can be expressed with higher probability. For illustration, we present an example of four artefacts, of which three are absolutely similar to each other and one is absolutely dissimilar to all others. All have equal probability.

$$\mathbf{K}^a = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{p}^a = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \tag{2}$$

The same information can be reduced to two absolutely dissimilar artefacts and the corresponding probabilities $\mathbf{p}^b$.

$$\mathbf{K}^b = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{p}^b = \begin{pmatrix} 0.25 \\ 0.75 \end{pmatrix} \tag{3}$$

Both representations yield the same VS, which reflects the imbalanced set of two absolutely dissimilar artefacts. $\mathrm{VS}(\mathbf{K}^a, \mathbf{p}^a) = \mathrm{VS}(\mathbf{K}^b, \mathbf{p}^b) = 1.755\ldots$

The imbalance of our example set negatively affects its diversity. If all items in the set are given equal importance, one artefact is under-represented. Instead, each of the two absolutely dissimilar artefacts in the set should thus be assigned equal weight $p = 0.5$. In the case of repetitions, this weight has to be divided across the repeated artefacts.

$$\mathbf{K}^c = \begin{pmatrix} 1&0&0&0 \\ 0&1&1&1 \\ 0&1&1&1 \\ 0&1&1&1 \end{pmatrix}, \quad \mathbf{p}^c = \begin{pmatrix} 0.5 \\ 0.166\ldots \\ 0.166\ldots \\ 0.166\ldots \end{pmatrix} \tag{4}$$

This maximises VS to reflect the effective number of absolutely dissimilar artefacts $\mathrm{VS}(\mathbf{K}^c, \mathbf{p}^c) = 2$.

Importance Sampling

Conventionally, training examples are drawn from a dataset with uniform probability. In importance sampling, instead, examples are chosen according to their contribution to an unknown target distribution. In our case, the importance of training examples is determined by their individual contribution to the overall dataset diversity as quantified by the optimised probability distribution p (see example above). We aim to increase the output diversity of a model. For this, we replace the basic sampling operation by a diversity-weighted importance sampling scheme.

Model Evaluation

To assess model performance, we use some common measures for generative models, as well as measures specifically relevant to our method. Inception Score (IS) (Salimans et al., 2016), Fréchet Inception Distance (FID) (Heusel et al., 2017), and Precision-Recall (k-NN parameter $k = 3$) (Kynkäänniemi et al., 2019) quantify sample fidelity and mode coverage w.r.t. the unbiased training data distribution. We employ our Weighted Fréchet Inception Distance (wFID) to account for the change in target distribution, induced by our method through diversity-weighted sampling (see below for details). Diversity is estimated with the Vendi Score (VS) (Friedman and Dieng, 2022).

Note that we follow the recommendations by Barratt and Sharma (2018) and calculate IS over the entire generated set of samples, removing the common split into subsets. We also remove the exponential, such that the score becomes interpretable in terms of mutual information. While not all reported scores are directly comparable to other works, our measurements are internally consistent and reliable.

Image Embeddings Instead of comparing image data on raw pixels, standard evaluation measures of model performance have relied on image classification networks to be used as embedding models for feature extraction. The InceptionV3 model (Szegedy et al., 2016) is most commonly used as a representative feature space and has been widely adopted as part of a standard measurement pipeline. Unfortunately, small numerical differences in model weights, implementations and interpolation operations can compound to bigger discrepancies. For example, image scaling to match the input size of an embedding model can change the computed features and thus affect the subsequent measurements (Parmar, Zhang, and Zhu, 2022). Furthermore, embedding models trained on the ImageNet dataset, like InceptionV3, inherit the dataset's biases, which can lead to unreliable measurements that do not agree with human assessment (Kynkäänniemi et al., 2023). In this work, we therefore follow the recommendations for anti-aliasing re-scaling and use CLIP ViT-L/14 (Radford et al., 2021) as the image embedding model in our feature extraction and measurement pipelines (except for IS). Note that, while trained on a much larger (proprietary) dataset and better suited as embedding model, CLIP still has its own biases.

Diversity Weights

If artefacts in a set are repeated, i.e. their relative abundance is increased, their individual contribution to the overall diversity of the set decreases. Yet, with uniform weighting, all artefacts contribute to the model distribution equally (cf. eq. 2). Instead, we aim to adjust the weight of individual artefacts in a set in accordance with their contribution to overall diversity.

We formulate an optimisation problem to find the optimal weight for each artefact in a set, such that its diversity, as measured by VS, is maximised.

$$\max \exp\Big(-\sum^{n} \lambda_i \log \lambda_i\Big) \tag{5}$$

$$\begin{aligned} \text{s.t. } & \sum^{n} p_i = 1 \qquad 0 \le p_i \le 1 \\ \text{where } & \mathbf{p} = (p_i, \ldots, p_n), \;\; p_i \in \mathbb{R}^{[0,1]} \\ & \mathbf{K} \in \mathbb{R}^{n \times n}, \;\; \mathbf{K}_{ii} = 1 \\ & \mathbf{K}^{\mathbf{p}} = \mathrm{diag}(\sqrt{\mathbf{p}})\, \mathbf{K}\, \mathrm{diag}(\sqrt{\mathbf{p}}) = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{-1} \\ & \boldsymbol{\lambda} = \mathrm{diag}(\boldsymbol{\Lambda}) = (\lambda_1, \ldots, \lambda_n) \end{aligned}$$

Optimisation Algorithm[1]

We compute an approximate solution to the optimisation problem via gradient descent (algorithm 1). The objective function consists of two terms: diversity loss and entropy loss. The diversity loss is defined as the negative probability-weighted VS of the set of artefacts, given its similarity matrix and the corresponding probability vector (cf. eq. 1). To ensure the optimised artefact probability distribution follows the Kolmogorov (1933) axioms, we make the following adjustments. Instead of optimising the artefact probabilities directly, we optimise a weight vector $\mathbf{w}$. The probability vector $\mathbf{p}$ is obtained by dividing the $\mathbf{w}$ by the sum of its values, which guarantees the second axiom. To satisfy the first axiom, we implement a fully differentiable version of VS in log space. Optimising in log space enforces weights above zero, since the logarithm $\log x$ is only defined for $x > 0$ and tends to negative infinity as x approaches zero. However, if the weights have no upper limit, values can grow

[1]An implementation of the optimisation algorithm is available at https://github.com/sebastianberns/diversity-weights

Algorithm 1 Vendi Score Diversity Weight Optimisation

Input: Similarity matrix $\mathbf{K}$ of N artefacts
Parameter: Loss term balance γ, num iterations I, learning rate α, Adam hyperparams β_1, β_2

1: Initialise $\mathbf{w} = (w_1, \ldots, w_N)$, where $w_i = 1$
2: **for** $i = 0$ **to** I **do**
3: $\quad \mathbf{p} \leftarrow \mathbf{w} / \sum w_i$
4: $\quad g \leftarrow -\nabla_{\mathbf{p}}\, \gamma \mathrm{VS}(\mathbf{K}, \mathbf{p}) - (\gamma - 1)\, \mathrm{H}(\mathbf{p})$
5: $\quad \mathbf{w} \leftarrow \mathrm{Adam}(\mathbf{w}, g, \alpha, \beta_1, \beta_2)$
6: **end for**

Output: Weight vector $\mathbf{w}$

unbounded. A heavy-tailed weight distribution negatively affects the importance sampling step of our method during training, as batches can become saturated with the highest-weighted training examples, causing overfitting. We therefore add an entropy loss term $\mathrm{H}(\mathbf{p}) = -\sum p_i \log(p_i)$ to be maximised in conjunction with the diversity loss. The entropy loss acts as a regularisation term over the weight vector, such that its distribution is kept as close to uniform as possible. The emphasis on the two loss terms is balanced by the hyperparameter $\gamma \in [0, 1]$.

$$\mathcal{L} = -\gamma \mathrm{VS}(\mathbf{K}, \mathbf{p}) - (\gamma - 1)\, \mathrm{H}(\mathbf{p}), \qquad \mathbf{p} = \frac{\mathbf{w}}{||\mathbf{w}||_1} \quad (6)$$

Given a normalised data matrix X where rows are examples and columns are features, we obtain the similarity matrix $\mathbf{K}$ by computing the Gram matrix $K = X \cdot X^{\mathsf{T}}$. The weight vector $\mathbf{w}$ is initialised with uniform weights $w_i = \log(1) = 0$. The probability vector $\mathbf{p}$ is obtained by dividing the weight vector $\mathbf{w}$ by the sum of its values. We choose the Adam optimiser (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate decays exponentially every 5 iterations by a factor of 0.99.

Weighted FID

The performance of generative models, in particular that of implicit models like GANs, is conventionally evaluated with the FID (Heusel et al., 2017). Raw pixel images are embedded into a representation space, typically of an artificial neural network. Assuming multi-variate normality of the embeddings, FID then estimates the distance between the model distribution and the data distributions from their sample means and covariance matrices.

In our proposed method, however, the learned distribution is modelled on a weighted version of the dataset. Moreover, referring to the standard statistics of the original dataset is no longer applicable, as the weighted sampling scheme changes the target distribution. We therefore adjust the measure such that it becomes the Weighted Fréchet Inception Distance (wFID), where the standard mean and covariances to calculate the dataset statistics are substituted by the weighted mean $\mu^* = \left(\sum w_i \mathbf{x}_i\right) / \sum w_i$ and the weighted sample covariance $\mathbf{C} = \left(\sum w_i (\mathbf{x}_i - \mu^*)^{\mathsf{T}} (\mathbf{x}_i - \mu^*)\right) / \sum w_i$. Note that the statistics of the model distribution need to be calculated without weights as the model should have learned the diversity-weighted target distribution.

Proof-Of-Concept Study on Hand-Written Digits

We show the effect of the proposed method in an illustrative study on pairs of handwritten digits. While artistically not particularly challenging, digit pairs have several benefits over other exemplary datasets. First, the pairings of digits create a controlled setting with two known types of artefacts. Second, hand-written digits present a simple modelling task, in which the quality and diversity of a model's output is easy to visually assess. And third, generating digits is fairly uncontroversial. While, for example, generating human faces is more relevant for the subject of diversity, it is also a highly complex and potentially emotive domain.

Methodology

For individual pairs of digits, we quantitatively and qualitatively evaluate the results of GAN training with diversity weights and compare it against standard training. Experiments are repeated five times with different random seeds.

Digit Pairs From the ten classes of the MNIST training set, we select three digit pairs: 0-1, 3-8, and 4-9, which represent examples of similar and dissimilar pairings. For example, images of hand-written zeros and ones are easy to distinguish, as they are either written as circles or straight lines. In contrast, threes and eights are both composed of similar circular elements.

Balanced Datasets For each pair of digits, we create five balanced datasets (with different random seeds) of 6,000 samples each. Each dataset consists of 3,000 samples of either digit, randomly selected from the MNIST training set. We compute features by embedding all images using the CLIP ViT-L/14 model. To optimise the corresponding diversity weights, we obtain pairwise similarities between images by calculating the Gram matrix of features.

Diversity Weights For each dataset (5 random draws per digit pair), we optimise the diversity weights for 100 iterations. We fine-tune the loss term balance hyperparameter and determine its optimal value $\gamma = 0.8$, where the weights converge to a stable distribution, while reaching a diversity loss as close to the maximum as possible. Without the entropy loss term ($\gamma = 1.0$) the weights yield the highest VS, but reach both very high and very low values. Large differences in weight values negatively affect the importance sampling step of our method during training, as batches can become saturated with the highest-weighted training examples. In contrast, a bigger emphasis on the entropy loss ($\gamma = 0.6$) results in the weights distribution being closer to uniform, but does not maximise diversity. The hyperparameter γ provides control over the trade-off between diversity and *typicality*, i.e. the extent to which an generated artefact is a typical training example (Ritchie, 2007). The VS of the digit datasets when measured without and with diversity weights at different loss term balances are presented in table 1.

Table 1: Vendi Score (VS) of digit pair datasets (mean ± std dev) with uniform and diversity weights with different loss balances γ

VS weights	MNIST digit pairs		
	Pair 0-1	Pair 3-8	Pair 4-9
Uniform weights	1.77±0.003	1.96±0.004	2.07±0.004
DivW ($\gamma = 0.6$)	2.13±0.020	2.64±0.016	2.65±0.010
DivW ($\gamma = 0.8$)	2.79±0.052	3.45±0.027	3.38±0.025
DivW ($\gamma = 1.0$)	3.08±0.046	3.67±0.023	3.60±0.023

The resulting diversity weight for each of the 6,000 samples corresponds to their individual contributions to the overall diversity of the dataset. We give an overview of the highest and lowest weighted data samples in fig. 2. Low-weighted samples are typical examples of the MNIST dataset: e.g. round zeros and simple straight ones, all of similar line width. High-weighted samples show a much greater diversity: thin and thick lines, imperfect circles as zeros, ones with nose and foot line.

Training For each digit dataset, we compare two training schemes: 1) a baseline model with the standard training scheme, and 2) three models trained with our diversity weights (DivW) method and different loss term balances (γ), where training examples are drawn according to the corresponding diversity weights. The compared loss term balances are $\gamma = 0.6$, $\gamma = 0.8$, and $\gamma = 1.0$. All models have identical architectures (Wasserstein GAN with gradient penalty; Gulrajani et al., 2017) and hyperparameters and are optimised for 6,000 steps (see appendix for details).

To allow our method to develop its full potential, we increase the batch size to 6,000 samples, the size of the dataset. Training examples are drawn according to diversity weights *with* replacement, i.e. the same example can be included in a batch more than once. Small batches in turn would be dominated by the highest-weighted examples, causing overfitting and ultimately mode collapse.

Evaluation We evaluate individual models on six measures: Vendi Score (VS) to quantify output diversity; Inception Score (IS), Fréchet Inception Distance (FID) and weighted FID (wFID), as well as Precision–Recall (PR) to estimate sample fidelity and mode coverage. From each model we obtain 6,000 random samples, the same amount as a digit dataset. As described above, for all measures, except IS, we use CLIP as the image embedding model to compute image features. For VS, we obtain pairwise similarities between images by calculating the Gram matrix of features. Our proposed wFID measure accounts for the different target distribution induced by the diversity weights.

Results

An overview of our quantitative results in given in fig. 3. For three pairs of digits, we compare our diversity weights (DivW) method with three different loss term balances (γ)

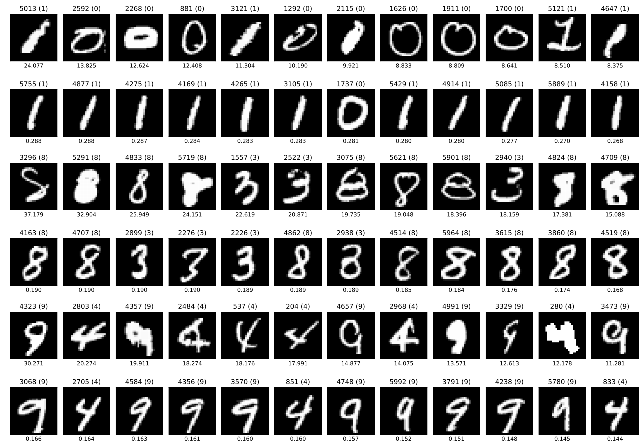


Figure 2: Digits ordered by diversity weight (index above with label in brackets, weight below). First two rows: pair 0-1, two middle rows: pair 3-8, last two rows: pair 4-9. Odd rows: twelve highest weighted, even row: twelve lowest weighted.

against a standard GAN. The balance of loss terms determines the emphasis on a uniform distribution of weights (lower γ) over higher diversity (higher γ). Accordingly, in the diversity weight optimisation, a balance of $\gamma = 1.0$ corresponds to a full emphasis on diversity and no entropy loss, while $\gamma = 0.5$ strikes an equal balance between the two.

Our results agree on almost all measures across all three digit pairs, except on IS which we discuss further below. As expected, the higher the emphasis on the diversity loss, the higher (and better) the VS (fig. 3, top left). This comes with a trade-off in sample fidelity and mode coverage, as quantified by PR (fig. 3, middle and bottom left) and FID (fig. 3, top right). However, when accounting for a weighted training dataset with our Weighted FID measure, the distance of our DivW model distribution to the target distribution is notably lower than or at least on par with the standard model (fig. 3, middle right).

Results on IS (fig. 3, bottom right) show the difficulty in distinguishing different pairs of digits. For the pairing 0-1, the standard model and the DivW $\gamma = 0.6$ model score notably higher than the other two DivW models ($\gamma = 0.8$ and $\gamma = 1.0$), while their scores are lower for the pairings 3-8 and 4-9. This suggests that, even for the standard model it is difficult to model two similar digits like 3-8 and 4-9.

For visual inspection and qualitative analysis, we provide random samples in fig. 4 for all digit pairs and models.

Discussion

In recent years, research communities have become better aware of data biases and their impact on society through the proliferation of data-driven technologies. Likewise, CC researchers have highlighted its potential implications for CC research and the importance of mitigation (Smith, 2017; Loughran, 2022). Real-world datasets are a limited sample of a complex world and should not be considered the 'ground truth', or as representing the 'true' distribution. This practical impossibility further motivates our proposal to

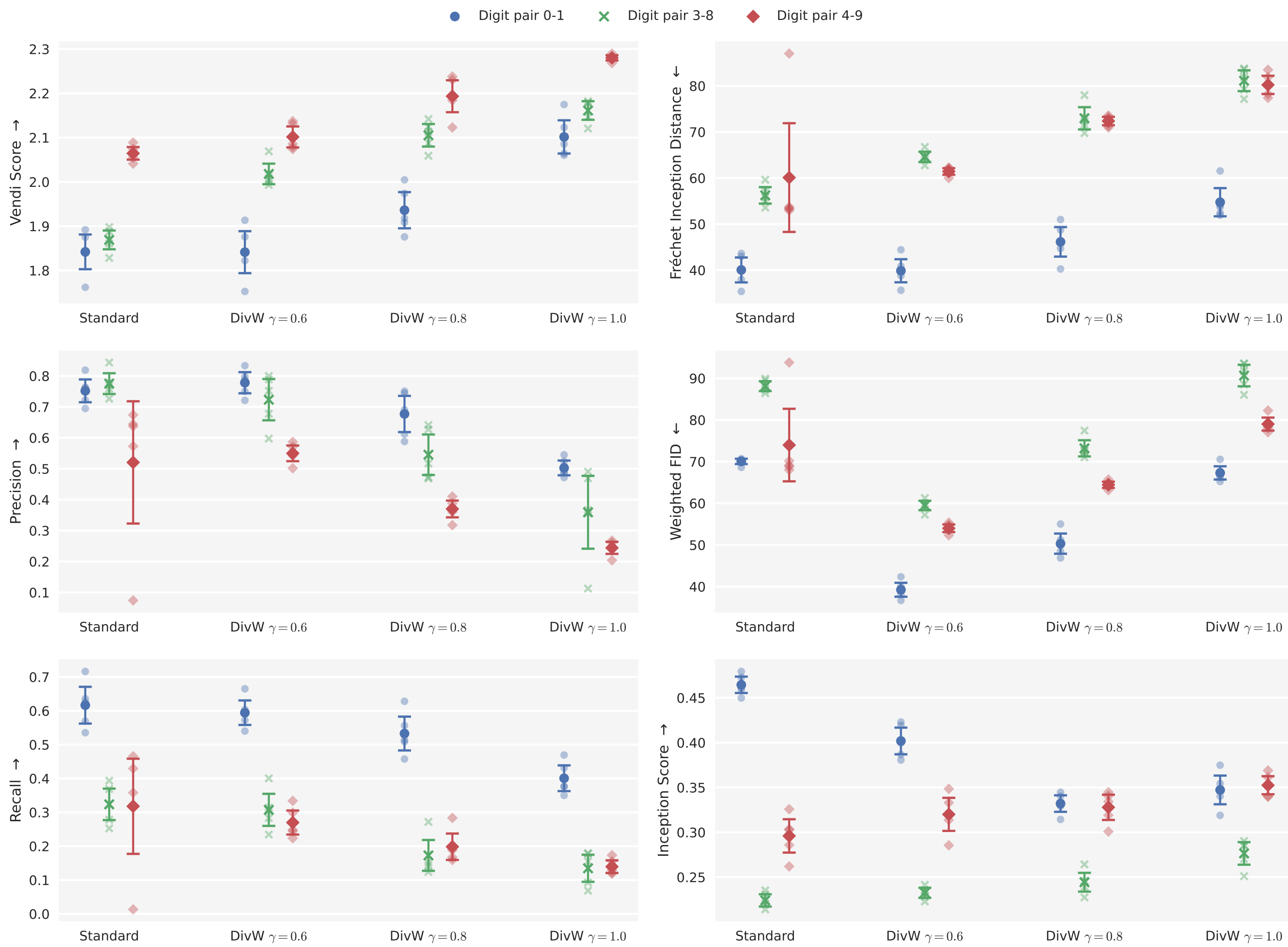


Figure 3: Performance comparison of our method (DivW) with different loss term balances (γ) against a standard GAN, trained on three digit pair datasets (blue circles: 0-1, green crosses: 3-8, red diamonds: 4-9) with six measures: VS, PR and IS (higher is better), as well as standard FID and weighted FID scores (lower is better). Means and 95 % confidence intervals over five random seeds. Individual datapoints show means over five random sampling repetitions. The hyperparameter γ provides control over the trade-off between diversity and typicality.

shift away from the predominant mode coverage paradigm.

Ongoing debates have not yet resulted in a uniformly accepted way of dealing with data bias in generative machine learning more generally, and CC specifically. One way to address data bias is to gather more or better data. But this is not always possible or practical, since collecting, curating and pre-processing new data is notoriously laborious, costly, or subject to limited access. Another way is to instead adjust the methodology of learning from data, such that a known data bias is mitigated. In this work, we focus on the latter and propose the *diversity-weighted sampling* scheme to address the imbalance of representation between majority and minority features in a dataset.

Diversity weights address the specific bias of *data imbalance*, particularly in unsupervised learning. In contrast to supervised settings, where class labels provide a clear categorisation of training examples, here common features are often shared between various types of examples. This makes it difficult to find an appropriate balance of training examples. Diversity weights give an indication of which type of examples are under-represented from a diversity-maximisation perspective. We draw a connection to issues of DEI as data biases often negatively affect under-represented groups (Bolukbasi et al., 2016; Zhao et al., 2017; Hendricks et al., 2018; Stock and Cisse, 2018).

Combining image generation models with multi-modal embedding models, like CLIP, enables complex text-to-image generative systems which can be doubly affected by data bias through the use of two data-driven models: the image generator and the image-text embedding. The discussion on embedding models, and other methods that can guide the search for artefacts, is beyond the scope of this paper. Our work focuses on the image generators powering these technologies. Yet, a conscious shift to *mode balancing*, in particular for the training of the underlying generative model, could support the mitigation of bias in text-to-image generation models, complementing existing efforts in prompt engineering after training (Colton, 2022).

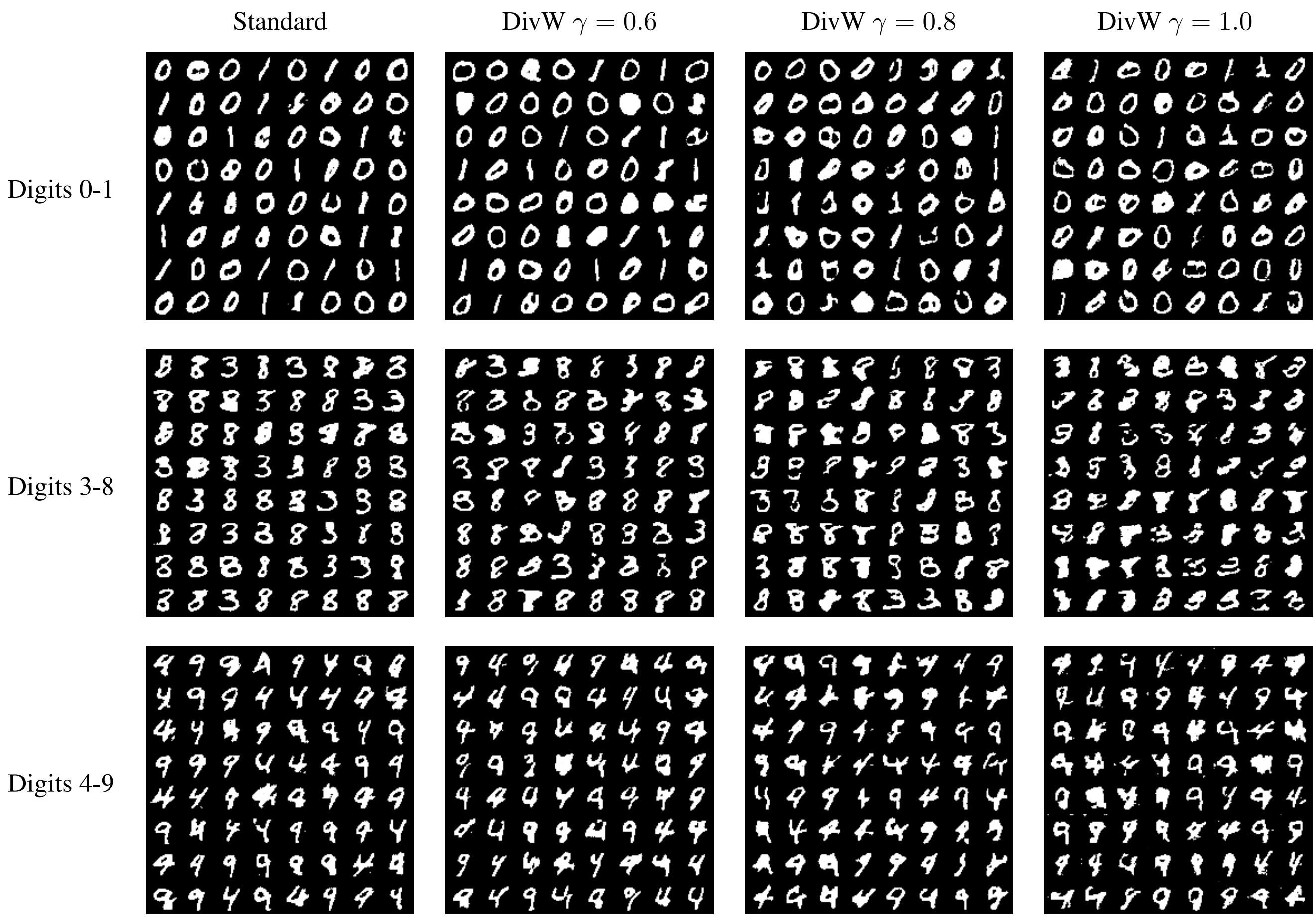


Figure 4: Random samples for all digit pairs (top row: 0-1, middle: 3-8, bottom: 4-9) from the standard models (left column) and our DivW models with different loss balances (γ). The hyperparameter γ provides control over the trade-off between diversity and typicality.

It is worth noting, that our method also introduces bias, particularly emphasising under-represented features in the dataset. We do this explicitly and for a specific purpose. Other applications might differ in their perspective and objective and deem none or other biases less or more important. As we mentioned above, since a dataset cannot maintain its status of 'ground truth', the responsibility of reviewing and potentially mitigating data biases falls onto researchers and practitioners. We hope our work proves helpful in this task.

Related Work

Previous work primarily focuses on samples from minority groups and related data biases. Objectives range from mitigating such biases to improving minority coverage, i.e. achieving better image fidelity for underrepresented data examples. Some approaches employ importance weighting where weights are derived from density ratios, either via an approximation based on the discriminator's prediction (Lee et al., 2021) or via an additional probabilistic classifier (Grover et al., 2019). Others propose an implicit maximum likelihood estimation framework to improve the overall mode coverage (Yu et al., 2020). These methods either depend on additional adversarily trained models or on more specific hybrid models. Our approach, instead, has two major benefits over previous work. First, it is model-agnostic and thus potentially applicable to a wide range of network architectures and training schemes. Second, it only adds an offline pre-computation step prior to conventional training procedures and during training solely intervenes at the data sampling stage.

Authors of previous work further argue for increased diversity, but do not evaluate on explicit measures of diversity. Results are reported on the standard metrics IS, as well as FID and PR which rely on the training dataset for reference. Consequently, they can only estimate sample fidelity and mode coverage as present in the data. We, instead, evaluate on measures designed to objectively quantify diversity.

Most importantly, while we argue for an adaptation of modelling techniques to allow for *mode balancing* to achieve higher output diversity, all related works operate under the *mode coverage* paradigm. In fact, Lee et al. (2021) include a discriminator rejection sampling step (Azadi et al., 2018) after training to undo the bias introduced by their importance sampling scheme.

Conclusions

We introduced a method to derive a weight vector over the examples in a training dataset, which indicate their individual contribution to the dataset's overall diversity. *Diversity weights* allow to train a generative model with importance sampling such that the model's output diversity increases.

Our work is motivated by potential benefits for computational creativity applications which aim to produce a wide range of diverse output for further design iterations, ranging from artistic over constrained to scientific creativity. We also highlight a connection to issues of data bias in generative machine learning, in particular data imbalances and the under-representation of minority features. The impracticality of easily mitigating data imbalances in an unsupervised setting further motivates our work.

In a proof-of-concept study, we demonstrated that our method increases model output diversity when compared to a standard GAN. The results highlight a trade-off between artefact typicality, i.e. the extent to which an artefact is a typical training example, and diversity. Our method provides control over this trade-off via a loss balance hyperparameter.

Future Work

We plan to build on the present work in several ways. First, by refining our method, in particular the training procedure, to improve overall sample fidelity. For this, a thorough analysis and systematic comparison to related work is needed. The loss balance hyperparameter could further be tuned automatically by including it as a learnable parameter in the optimisation procedure. Apart from our gradient descent approach, there might be alternative exact or approximate methods for the diversity weight optimisation, e.g. constraint optimisation or analytical solutions.

Second, we plan to extend experimentation to other generative models and on bigger and more complex datasets to demonstrate the scalability of our approach. Since our method is architecture-agnostic, there remain many opportunities for future work to understand the effect and potential benefits of our method in other modelling techniques. As GAN training is notoriously unstable and requires careful tuning, other modelling techniques might prove more appropriate. Results on datasets representing humans are needed to demonstrate the capability of our method to mitigate issues of DEI resulting from data imbalances.

Moreover, empirical studies will be necessary to investigate how the shift from mode coverage to mode balancing can support diversity in a large range of CC applications.

Acknowledgements

Sebastian Berns is funded by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) [EP/S022325/1]. Experiments were performed on the Queen Mary University of London Apocrita HPC facility, supported by QMUL Research-IT (King, Butcher, and Zalewski, 2017).

Appendix

The tables below outline the experiments' training hyperparameters and network architectures, which do not include any pooling, batchnorm or dropout layers. He initialisation (Kaiming uniform) is used for convolutional layers (conventional and upsampling) and Glorot initialisation (Xavier uniform) for fully connected (FC) layers.

Table 2: Architecture of WGAN-GP generator network. Upsampling convolutional layers (ConvTranspose) have kernel size 4×4, stride 2, padding 1, dilation 1.

WGAN Generator		
Layer	**Output**	**Activation**
Input z	64	
Linear (FC)	2,048	ReLU
Reshape	$4 \times 4 \times 128$	
ConvTranspose	$8 \times 8 \times 64$	ReLU
Cut	$7 \times 7 \times 64$	
ConvTranspose	$14 \times 14 \times 32$	ReLU
ConvTranspose	$28 \times 28 \times 1$	Sigmoid

Table 3: Architecture of WGAN-GP critic network. Convolutional layers have kernel size 5×5, stride 2, padding 2.

WGAN Critic		
Layer	**Output**	**Activation**
Input	$28 \times 28 \times 1$	
Conv	$14 \times 14 \times 32$	LeakyReLU(0.2)
Conv	$7 \times 7 \times 64$	LeakyReLU(0.2)
Conv	$4 \times 4 \times 128$	LeakyReLU(0.2)
Reshape	2,048	
Linear (FC)	1	

Table 4: Training hyperparameters

Hyperparameter	**Value**
Num steps	6,000
Num critic steps	5
Batch size	6,000
GP weight	10.0
LR generator	0.0001
LR critic	0.0001
Adam β_1	0.5
Adam β_2	0.9

References

Azadi, S.; Olsson, C.; Darrell, T.; Goodfellow, I.; and Odena, A. 2018. Discriminator Rejection Sampling. In *ICLR*.

Barratt, S., and Sharma, R. 2018. A Note on the Inception Score. In *Proceedings of the ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*.

Berns, S., and Colton, S. 2020. Bridging Generative Deep Learning and Computational Creativity. In *Proceedings of ICCC*.

Berns, S.; Broad, T.; Guckelsberger, C.; and Colton, S. 2021. Automating Generative Deep Learning for Artistic Purposes: Challenges and Opportunities. In *Proceedings of ICCC*.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge, second edition.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in NeurIPS*, volume 29.

Bradner, E.; Iorio, F.; and Davis, M. 2014. Parameters Tell the Design Story: Ideation and Abstraction in Design Optimization. In *Proceedings of the Symposium on Simulation for Architecture & Urban Design*.

Broad, T.; Berns, S.; Colton, S.; and Grierson, M. 2021. Active Divergence with Generative Deep Learning - A Survey and Taxonomy. In *Proceedings of ICCC*.

Colton, S. 2022. Towards Educating Artificial Neural Systems. In *Proceedings of the International Workshop on Neuro-Symbolic Learning and Reasoning*.

Friedman, D., and Dieng, A. B. 2022. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. ArXiv:2210.02410v1.

Grover, A.; Song, J.; Kapoor, A.; Tran, K.; Agarwal, A.; Horvitz, E. J.; and Ermon, S. 2019. Bias Correction of Learned Generative Models using Likelihood-Free Importance Weighting. In *Advances in NeurIPS*, volume 32.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved Training of Wasserstein GANs. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in NeurIPS*, volume 30.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of ECCV*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in NeurIPS*, volume 30.

Kandpal, N.; Wallace, E.; and Raffel, C. 2022. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In *Proceedings of ICML*.

King, T.; Butcher, S.; and Zalewski, L. 2017. Apocrita - High Performance Computing Cluster for Queen Mary University of London. DOI:10.5281/zenodo.438045.

Kingma, D. P., and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Ko, H.-K.; Park, G.; Jeon, H.; Jo, J.; Kim, J.; and Seo, J. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *International Conference on Intelligent User Interfaces*.

Kolmogorov, A. 1933. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer-Verlag.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in NeurIPS*.

Kynkäänniemi, T.; Karras, T.; Aittala, M.; Aila, T.; and Lehtinen, J. 2023. The Role of ImageNet Classes in Fréchet Inception Distance. In *ICLR*.

Lee, J.; Hong, Y.; Kim, H.; and Chung, H. W. 2021. Self-Diagnosing GAN: Diagnosing Underrepresented Samples in Generative Adversarial Networks. In *Advances in NeurIPS*, volume 34.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2014. Computational Game Creativity. In *Proceedings of ICCC*.

Loughran, R., and O'Neill, M. 2017. Application Domains Considered in Computational Creativity. In *Proceedings of ICCC*.

Loughran, R. 2022. Bias and Creativity. In *Proceedings of ICCC*.

Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; and Naik, N. 2023. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of ICML*.

Razeghi, Y.; Logan IV, R. L.; Gardner, M.; and Singh, S. 2022. Impact of Pretraining Term Frequencies on Few-Shot Reasoning. ArXiv:2202.07206v2.

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines* 17(1).

Runco, M. A., and Jaeger, G. J. 2012. The Standard Definition of Creativity. *Creativity Research Journal* 24(1).

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Gontijo-Lopes, R.; Ayan, B. K.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in NeurIPS*, volume 35.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Advances in NeurIPS*, volume 29.

Smith, G. 2017. Computational Creativity and Social Justice: Defining the Intellectual Landscape. In *Proceedings of the Workshop on Computational Creativity and Social Justice at ICCC*.

Stanley, K. O., and Lehman, J. 2015. *Why Greatness Cannot Be Planned: The Myth of the Objective*. Springer.

Stock, P., and Cisse, M. 2018. ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. In *Proceedings of ECCV*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of CVPR*.

Veale, T.; Cardoso, F. A.; and Pérez y Pérez, R. 2019. Systematizing Creativity: A Computational View. *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems* 1–19.

Vimpari, V.; Kultima, A.; Hämäläinen, P.; and Guckelsberger, C. 2023. “An Adapt-or-Die Type of Situation”: Perception, Adoption, and Use of Text-To-Image-Generation AI by Game Industry Professionals. ArXiv:2302.12601v3.

Volz, V.; Schrum, J.; Liu, J.; Lucas, S. M.; Smith, A.; and Risi, S. 2018. Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network. In *Proceedings of GECCO*.

Yu, N.; Li, K.; Zhou, P.; Malik, J.; Davis, L.; and Fritz, M. 2020. Inclusive GAN: Improving Data and Minority Coverage in Generative Models. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Proceedings of ECCV*. Springer International Publishing.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Zhong, P.; Mo, Y.; Xiao, C.; Chen, P.; and Zheng, C. 2019. Rethinking Generative Mode Coverage: A Pointwise Guaranteed Approach. In *Advances in NeurIPS*, volume 32.

Climate Implications of Diffusion-based Generative Visual AI Systems and their Mass Adoption

Vanessa Utz and Steve DiPaola

School of Interactive Arts and Technology (SIAT)
Simon Fraser University
Burnaby, BC V5A 1S6 Canada
vutz@sfu.ca sdipaola@sfu.ca

Abstract

Climate implications of rapidly developing digital technologies, such as blockchains and the associated crypto mining and NFT minting, have been well documented and their massive GPU energy use has been identified as a cause for concern. However, we postulate that due to their more mainstream consumer appeal, the GPU use of text-prompt based diffusion AI art systems also requires thoughtful considerations. Given the recent explosion in the number of highly sophisticated generative art systems and their rapid adoption by consumers and creative professionals, the impact of these systems on the climate needs to be carefully considered. In this work, we report on the growth of diffusion-based visual AI systems, their patterns of use, growth and the implications on the climate. Our estimates show that the mass adoption of these tools potentially contributes considerably to global energy consumption. We end this paper with our thoughts on solutions and future areas of inquiry as well as associated difficulties, including the lack of publicly available data.

Introduction

Many of today's rapidly developing digital technologies are being critically discussed due to their potential climate impact. For instance, blockchain-based technologies and the associated crypto-mining and NFT (Non-Fungible Tokens) minting, have frequently come under fire for their high energy usage (deVries 2019). While AI has often been proposed as a solution to the problems digital technologies create for our environment, interest has now also shifted to machine learning (ML) and its environmental implications. However, so far most of the work in this area has focused on the training phase of large AI systems. Our position is that more effort needs to be put into the study of the environmental impact that AI systems have during their usage (referred to technically as inference), particularly given the recent explosion in commercially available AI art systems that have been shown to have large scale consumer appeal.

While the environmental consequences that are associated with the energy consumption required to power AI tools, are not unique to the creative AI (cAI) field (including Computational Creativity), we want to raise awareness of this issue among cAI researchers and encourage further scientific investigations. It is also our hope to illustrate that this issue can be positioned as a user education concern.

We begin this position paper by providing some foundational background information into the discourse surrounding technology and its climate impact, both good and bad. We then discuss the current state of research into the energy consumption of ML systems. We also provide an overview of popular AI systems for visual art creation and provide data on their scale. To support our position regarding the need for more research into the environmental impact of cAI systems, we then proceed to demonstrate a preliminary analysis of the energy consumption of generative visual AI systems and compare the data to the climate impact of other digital technologies. We end this paper by drawing a connection to contemporary discussions on waste and overconsumption in the digital space. We also highlight future research avenues such as inquiries into the psychological principles that lead to the prolonged use of these tools.

Technology and Climate

We begin this topic with a brief introduction to the discourse on technology and climate. For this purpose, we will begin talking about technology and climate in general terms, before focusing on our main concern around energy consumption and carbon emissions.

Much of the conversation that views technology in a critical light appears to be focused on *electronic waste* (or *e-waste*). E-waste refers to the waste associated with the discarding of electronic devices that have reached their end of life (EOL) and has long been a focal point in discussions on pollution and the climate, with the Basel Convention labeling e-waste as hazardous over a decade ago (Widmer, Oswald-Krapf, Sinha-Khetriwal, Schnellmann and Böni 2005).

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

The main issues surrounding this debate involve the planned obsolescence of consumer electronics (Bisshop, Hendlin and Jaspers 2022), the low rate of recycled electronic devices/components (Perkins, Drisse, Nxele and Sly 2014), greenhouse gas (GHG) emissions (Singh and Ogunseitan 2022) and the pollutants contained within the devices, which have been shown to have detrimental health impacts on individuals exposed during the recycling/disposing process (Chen, Dietrich, Huo and Ho 2011).

Energy Consumption

Apart from the e–waste that is produced through modern day electronics, another concern involving technology and the climate, centers on the energy consumption of electronic devices while they are in use. Particularly with the rise in Blockchain technology much of the contemporary discourse on technology and climate has shifted to the immense energy consumption that is associated with these new technologies. Energy consumption is of concern here as increased energy consumption is linked to increases in GHG emissions (Luccioni and Hernandez-Garcia 2023; Schwartz, Dodge, Smith and Etzioni 2019).

Blockchain and the Internet of Things

Blockchain technology and the associated mining of cryptocurrencies and minting of NFTs has made headlines across the world over the last few years. In 2018 alone, the Bitcoin mining network was estimated to have consumed between 40 and 62.3 TWh of energy (which has been compared to the electricity consumption of major European countries like Hungary and Switzerland) (deVries 2019). According to the Bitcoin Energy Consumption Index (2023), in the first half of 2022, energy consumption of the Bitcoin network peaked at around 204.5 TWh per year. This peak lasted for approximately the first 6 months of 2022, before sharply declining to 91.31 TWh per year as of February 2023.

Alarm bells have also been going off in the area of the Internet of Things (IoT) and technological proliferation within our homes and environments (smart homes turning into smart cities). The need for investigations into the energy consumption of these systems has been made clear (Moutaib, Fattah and Farhaoui 2020).

Artificial Intelligence, the Savior?

In much of the literature discussed above, AI is often hailed as the solution to all our problems (for a brief literature review see: Dwivedi et al. 2022). Discussions involve incorporating AI into evermore technologies, such as Blockchain-based technologies, smart tech and data management systems, to increase efficiency and lower energy consumption and carbon emissions. Additionally, AI models are continuously proposed to aid in the mitigation of the impact of climate change. For instance, AI-based solutions have been suggested for aiding the food and agriculture sector (Ayed and Hanana 2021) and aiding with climate modeling (Huntingford, Jeffers, Bonsall, Christensen, Lees and Yang 2019).

Although these proposals are commendable, the impact of AI systems on the climate is rarely discussed in this type of literature. Even when these impacts are mentioned, they oftentimes are quickly glossed over instead of outlined thoughtfully and in much-needed detail (for example: PwC & Microsoft report on "How AI can enable a Sustainable Future" by Jobba and Herweijer n.d.).

Machine Learning and Energy Consumption

Although missing from much of the climate change focused AI literature, there is work on the environmental implications of ML that has been picking up momentum over the last few years. Energy consumption is a big focus here due to its relation to increased carbon emissions. Research in this area has focused mainly on the training of ML models. Here we will provide a quick overview of some of the work that has been taking place.

Strubell, Ganesh and McCallum (2019) took a closer look at deep learning models for NLP (Natural Learning Processing) and their financial and environmental costs. They posit that as models are becoming increasingly complex, more computational power is required leading to the increased use of powerful GPUs (Graphical Processing Units) and TPUs (Tensor Processing Units). Their study shows that training a model on a GPU has a similar amount of carbon emission as a trans-America flight. Lacoste, Luccioni, Schmidt and Dandres (2019) presented their Machine Learning Emissions Calculator with the goal that it would prove to be a useful tool for the ML community to track and estimate the carbon emissions during the training phase of ML models. More recently, Luccioni and Hernandez-Garcia (2023) published a survey of 95 ML models used in NLP and computer vision tasks. The survey contains data on energy consumption, CO2 emissions and how emissions volume relate to model performance. The results showed significant carbon emissions associated with the models that were reviewed and the authors call for a better understanding of the environmental impact of ML models within the community of researchers and developers.

The Mass Adoption of Creative AI

In the last few years, major developments have taken place in the cAI space, particularly with the rapid improvements that were seen in diffusion models. Not only are these models being used and tested among ML researchers and developers, but they have also shown to have mainstream consumer appeal and a large number of free as well as paid services and tools have now been developed and made available to the public. Some of the larger systems include DallE-2 (developed by OpenAI), Midjourney, Stable Diffusion (developed by Stability AI) and Artbreeder. It has been estimated that these four services alone produce over 20 million images per day (Note: this number is frequently cited online, however we were unable to independently corroborate this number) (Kelly 2022; Pennington 2022). The consumer appeal of these systems is also easily demonstrated by taking a closer look at the popularity of related mobile applications. For instance, when Google announced their Google Play's Best 2022 awards in November of said year, the best overall app was awarded to Dream by WOMBO (Lim 2022), a diffusion AI art generator that was released on November 16, 2021 (Wombo.ai 2021) and which as of late February 2023 has over 10 million downloads in the Google Play store. Although on WOMBO's website, the company indicates that globally the app has been downloaded over 100 million times, with a total image output of over 750 million (Wombo Inc. n.d.) Shortly afterwards, Lensa AI by Prisma Labs Inc., which was first released in 2018, exploded in popularity in December of 2022, when it reached over 12 million global downloads (Perez 2022), with 5.8 million downloads occurring in the first week of December alone (Ceci 2023a). The sharp increase in user numbers has been associated with the release of the app's new feature "Magic Avatars", which turns portraits into stylized imagery using a diffusion model. During the first weeks of December 2022, users spent approximately USD 9.25 million on the app's subscription and premium features according to data published by Statista (Ceci 2023b).

We will spend the remainder of this section introducing the most influential systems and highlighting their explosive growth over the last year. The purpose of this is to demonstrate the enormous computing power that is required to meet consumer demand. See Table 1 for a summary of user and daily image output numbers.

System/Platform	Total Users (in million)	Daily Image Output (in million)
Dall-E2	3	4
Midjourney	12[1]	Unknown[3]
Stable Diffusion	10	Unknown[3]
DreamStudio	1.5	3[4]
Dream	10-100[2]	1.6[4]
LensaAI	12[2]	Unknown[3]

Table 1 - Summary of user numbers and daily image output of a selection of the most popular generative AI art systems. [1]Number estimated based on members on official server (no numbers available for private servers), [2]Number estimated based on app downloads (might not be reflective of daily active users), [3]No estimates available based on limited data, [4]Number estimated based total image output since release date divided by days the system has been available.

Dall-E & Dall-E2

Dall-E was first released in January 2021, with the follow-up version Dall-E2 released in April 2022. The system was developed by OpenAI. The exact training data set has not been released by OpenAI. Dall-E2 is able to produce text-to-image and image-to-image output and is able to modify existing images (i.e. "inpainting"). Initially, the system was available through an invite-only service but has since been made available to a broader audience. According to OpenAI, as of November 2022, around 3 million people were using Dall-E2 to generate more than 4 million images per day (OpenAI 2022). OpenAI has also recently released an API (Application Programming Interface) which now enables developers to integrate Dall-E2 into their own applications, making the system even more widely available.

Midjourney

Developed by a research lab with the same name, this system was first released in April 2022. The newest version was released to the public in November 2022. Like OpenAI, Midjournery has not made their training data public. Midjournery operates as a bot through the Discord platform (an online communication platform which is divided into smaller communities, or so-called "servers", and allows text, voice and video chat). While the model as not been released to the public and the codebase and architecture are therefore unknown, according to StabilityAI's CEO Emad Mostaque, Midjourney has been leaning on Stable Diffusion since its beta lease (Mostaque 2022). As of February 2023, we have confirmed that the official Midjourney Discord server has over 12 million paying members (standard subscriptions currently start at USD30 per month). Members on the server are able to utilize the bot by providing text prompts to trigger image generation. It is important to note that on August 2nd, 2022, Midjourney announced on Twitter that the bot could be added to private servers and that users no longer had to use the official Midjourney server to generate images. We reached out to Midjourney but were unable to confirm how many servers the bot is currently operating.

Stable Diffusion (& DreamStudio)

Stable Diffusion uses a Latent Diffusion Model (LDM). It was trained on 2.3 billion images, contained within three datasets provided by LAION (Large-scale Artificial Intelligence Open Network): LAION-2B-EN, LAION-High-Resolution and LAION-Aesthetics 2v 5+. The initial release of Stable Diffusion took place in August 2022, with the stable release following in December 2022. The system was developed by StabilityAI. During an interview with Bloomberg in October 2022, StabilityAI's CEO confirmed that Stable Diffusion had over 10 million daily users and that their paid service DreamStudio has around 1.5 million active users (Fatunde and Tse 2022). Users on DreamStudio had generated a total of over 170 million images between launch and October 17th, 2022 (a time frame of 56 days) (StabilityAI 2022). StabilityAI has also noted that over 200k developers had downloaded their model (StabilityAI 2022). It is important to note that Stable Diffusion released their code and model to the public, therefore individuals are able to run the system locally on their own machines. Since the model is freely available and only requires under 10 GB of VRAM (video RAM) on widely available consumer GPUs (StabilityAI n.d.), it is difficult to estimate the true number of daily users. The open-source nature of Stable Diffusion has also caused it to be very commonly used as the system that powers many of the mobile AI art apps, such as Lensa AI (Hatmaker 2022), further increasing the difficulty associated with estimating its true reach and daily users.

The Energy Consumption of Generate AI Art Systems

Based on the numbers of daily users and daily output generation outlined in the previous section, we would like to provide some preliminary calculations into the energy consumption of these systems. It is important to point out that these numbers are at best a vast underestimation of how much energy is actually consumed, since we do not have access to the precise data on the usage of these systems.

We have decided to focus our initial calculations on Stable Diffusion, since proprietary information on, for instance, which data centers are used by Midjourney, introduces additional variables and unknowns into these calculations. We aim to provide a simple initial exploration into this topic.

Assumption 1: Hardware

According to sources, Stable Diffusion only natively supported NVIDIA RTX GPUs as of December 2022 (although it can be run in limited ways using other GPUs and CPUs) (Lewis 2022). However, according to StabilityAI's FAQ as of February 2023, most NVIDIA GPUs with 6GB or more, and high-end AMD GPUs are supported (n.d.). Additionally, NVIDIA RTX GPUs (particularly the RTX 3090) outperform most other commercially available GPUs in benchmark testing involving Stable Diffusion (Walton 2023). We therefore assume that this hardware is a reasonable scenario for energy consumption calculations. According to NVIDIA the Total Graphics Power (TGP) of the RTX 3090 is 350W, representing peak power draw (Burnes 2022; Cervenka 2022). We independently confirmed this by generating images using Stable Diffusion on a RTX 3090. Energy draw peaked at 350W when we generated a 1024x1024 image using the default 50 steps.

Assumption 2: Duration of Use

Regarding the duration of use of their hardware, we assume that the average user runs Stable Diffusion requiring peak power draw for approximately 1.5 hours per day. This assumption is based on a survey which we created to collect preliminary quantitative and qualitative data on the typical use of these systems. The survey was posted in seven AI art communities on Facebook and was live for 6 days, during which we collected 42 responses. The survey consisted of 5 multiple choice questions on the motivation behind their typical art creation, purpose of the final output/artwork, the type of system/tool used, their estimated total weekly image output, and their estimated average iteration per final artwork. The survey ended with one open-ended question asking participants to elaborate on their post-processing procedures, reuse and storing of images or any other information they deemed relevant to their creation process. On the question regarding their average weekly image generation, participants most frequently responded that they create over 1000 images per week. Unfortunately, we were unable to collect more specific data on the average output since we frankly did not expect such a high number of image generations for average users and limited our question to a maximum of 1000 images per week. However, during the open-ended section of the survey, a large portion of our respondents indicated they produce hundreds (sometimes thousands) of images in a single day using automated scripts. We therefore assume an average output of approximately 2000 images per week (or approximately 285 images per day). With this broad estimation we are trying to accommodate casual users as well as power-users. These calculations should be updated once better data becomes available. Based on our own testing on a RTX 3090, creating a 1024x1024 image using the default 50 steps, results in a generation time of 20 seconds. Based on this data, we assume that a user can generate up to 180 images per hour on a RTX 3090. To generate the target daily output of 285 images, a user needs to run the system at peak power draw for approximately 95 minutes per day.

Energy Consumption Calculations

Total Yearly Energy Consumption

Daily energy use per user:

$$350\ \text{W} \times 1.5\ \text{h} = 525\ \text{Wh} = 0.525\ \text{kWh} \tag{1}$$

Daily energy use for 10 million users:

$$0.525\ \text{kWh} \times 10{,}000{,}000 = 5{,}250{,}000\ \text{kWh} \tag{2}$$

Yearly energy use for 10 million users:

$$\begin{aligned} 365 \times 5{,}250{,}000\ \text{kWh} &= 1{,}916{,}250{,}000\ \text{kWh} \\ &= 1.92\ \text{TWh} \end{aligned} \tag{3}$$

Based on our assumption that the 10 million users of Stable Diffusion (as confirmed by StabilityAI) run the system for approximately 1.5 hours per day on a RTX 3090, will lead to a yearly energy consumption of approximately 1.92 TWh. This level of energy consumption is similar to the total electricity consumption of the West African nation Mauritania in 2021, which has been estimated to be 1.9 TWh according to the US Energy Information Administration (eia.gov n.d.).

Energy Consumption per Image

Number of images generated per hour:

$$3600 \text{ s per hour} \div 20 \text{ s required per images} = 180 \text{ images} \quad (4)$$

Energy use per image:

$$350 \text{ Wh} \div 180 \text{ images} = 1.94 \text{ Wh} \quad (5)$$

Extrapolation to Other Systems

In the following section, we are aiming to extrapolate the above data to a larger set of popular generative AI art systems that are either using Stable Diffusion code as their base (such as LensaAI) or are using similar diffusion-based technology. We estimate that the five popular systems Stable Diffusion (including DreamStudio), Midjourney, DallE-2, LensaAI, and Dream have approximately 48.5 million users (see Table 1 for a summary of user data).

Daily energy use per user:

$$350 \text{ W} \times 1.5 \text{ h} = 525 \text{ Wh} = 0.525 \text{ kWh} \quad (6)$$

Daily energy use for 48.5 million users:

$$0.525 \text{ kWh} \times 48{,}500{,}000 = 25{,}462{,}500 \text{ kWh} \quad (7)$$

Yearly energy use for 48.5 million users:

$$365 \times 25{,}462{,}500 \text{ kWh} = 9{,}293{,}812{,}500 \text{ kWh} = 9.29 \text{ TWh} \quad (8)$$

To put this estimation into perspective, based on numbers published by the US Energy Information Administration (eia.gov n.d.), the total electricity consumption of Kenya in 2021 was 9.1 TWh.

There are some obvious limitations with these numbers that need to be addressed. 1) Total users/downloads versus daily active users: the current calculations are based mainly on the publicly available data regarding app downloads, server members etc. This obviously differs from the actual number of daily users, since not every person who signs up for a subscription, downloads an app or installs a model on their device, is using the system regularly. For most of these systems (Stable Diffusion being the exception), we are unable to determine how many individuals use them on a daily basis. Making this data available to researchers and the general public would be a great step to increase transparency in this space. 2) Home computers versus cloud computing: our calculations are based on the peak energy draw of a commercially available GPU (RTX 3090) that has been shown to outperform many other processing units in benchmark testing. Apart from the fact that not all users own this specific GPU, we have no insight into how services like Midjourney generate their output. While we reached out to Midjourney, we were unable to receive a clear answer.

Comparison to Other Digital Technologies

So how does this compare to other digital technologies, such as blockchain related technologies and the training of AI systems, which are already being discussed as potentially being harmful due to their immense energy consumption?

According to the Bitcoin Energy Consumption Index (2023), current energy consumption sits at around 91.31 TWh per year. The energy consumption during the minting of an NFT (on the Ethereum blockchain) has been estimated to be approximately 142 kWh (Kshetri and Voas 2022) (this estimation was based on the energy use of the Ethereum blockchain before *The Merge*, a term referring to the moment on September 15, 2022 when the Ethereum blockchain moved from proof of work to proof of stake; a decision that lowered the blockchain's total energy demand by as much as 99.9996% (deVries 2022)).

Based on these numbers, it becomes evident that our estimation of the energy consumption of running Stable Diffusion on a home computer is vastly smaller than the total energy consumption associated with blockchain-based technology. However, expanding our estimation to include other systems, the total energy consumption becomes considerably more concerning. We would also like to 1) reiterate that this estimation is most likely a vast underestimation of the actual energy consumption involved in the entire AI art domain and 2) highlight how important an early call to action is, considering the potential and vast application spaces of these new tools. In our survey, 75.61% of participants responded that they would move from image generation to video generation once the tools become more widely available and easier to use. While we cannot put a concrete number on how this would affect energy consumption, this move would likely increase the need for more computing power and prolong the time for which these systems are in use. It is also important to note that video is not the only expansion on the horizon: animation, character bots, 3D gaming, and Virtual Reality (VR) environments are also being worked on in the generative AI space. For instance, StabilityAI recently partnered with KrikeyAI to develop text-to-animation tools (PRNewswire 2023; StabilityAI 2023).

Digital Waste

Now, we would like to take a step back and shift our conversation to the topic of digital waste and how these AI art systems have the potential to contribute to this type of waste. *Digital waste* (also referred to as *data waste* (Bietti and Vatanparast 2020)) is defined as "the carbon emissions, natural resource extraction, production of waste, and other harmful environmental impacts directly or indirectly attributable to data-driven infrastructures" (Bietti and

Vatanparast 2020 p.2). It is our position that the large-scale adoption of generative AI art tools may contribute to the replication of our modern society's overconsumption habits of natural resources within the digital space. What we are particularly referring to is the overconsumption of the generative tools themselves, and thereby producing large amounts of data that is not only energy-intensive to generate, but also subsequently needs to be stored and maintained in data centers.

To illustrate our concerns, we collected data on how users of generative AI tools interact with these systems and utilize the produced output. We used two data sources for this purpose: a series of polls that were posted by David Holz, founder of Midjourney, posted on the main Midjourney server on various days of January 2023, as well as the previously mentioned survey which we created (Note: the polls that were posted on the Midjourney server never close, so data from these polls might change in the future. The numbers we report were accessed on February 23rd, 2023). When analyzing our data, we identified two areas of interest which we have selected for further discussion: the motivation for using these tools and the utilization of the tool and output.

Utilization

Our survey has shown that the majority of users (57.14%) use a paid cloud service such as Midjourney, while 40.48% run a generative Art system on their home computers (the remainder of respondents use free cloud services). Most frequently, respondents claimed to create over 1000 images per week, with only 19.05% of respondents generating less than 100 images. We also identified power-users, who produce significantly larger image outputs, with one of our respondents explaining in the open-ended section of the survey that they "make 5000+ per night". The open-ended questions also showed that many users have automated the prompt-generation and use scripts to run the tools autonomously and continuously for hours. Finally, our data also showed that users frequently iterate on a single idea/prompt in order to get what they characterize as a successful piece. Approximately half of our respondents indicated that they require over 50 iterations on an idea to achieve a satisfying result, with 1 of our respondents regularly requiring over 500 interactions on a single idea.

Motivation

Participants in our survey as well as the official polls on Midjourney's server, show reliably that most users of these tools mainly create output for themselves. Our survey indicated that 38.1% of respondents use the tools solely for themselves as entertainment with occasionally sharing creations online. While a further 21.43% of our respondents indicated that their main motivation is related to sharing their creations with others. This number varies from the data shown in Midjourney's official polls where 98% of respondents (N = 568) indicated that they never shared any of their creations with others and are only creating them for themselves. We hypothesize that this difference could be attributed to our sampling strategy. While the Midjourney polls reached all users of the service, our survey was targeting users that were actively engaged in AI art focused social media communities. Both data sources also show that the proportion of professional users remains a minority, with our survey indicating that only 14.29% of respondents are professional users while the Midjourney polls (N = 3,203) show that 35.28% of respondents had used their generations within the context of their profession.

Digital Overconsumption

The data we have collected provides preliminary insight into how generative art systems are being deployed by users. We want to draw our reader's attention to the large amounts of data that are being generated and stored, never to be shared and consumed, and without an obvious utility. We want to lean on work by Brown & Cameron (2000), which categorizes overconsumption as a form of the common pool source dilemma, 1) where the size of the resource pool is not known, 2) access to the resources in not equally distributed among individuals and 3) individuals must make decision on their consumption of goods and services without a full picture of the quantities and types of resources required in the process. We believe that this dilemma applies here as well, as a significant portion of users are likely unaware of the total global energy consumption (and the associated GHG emissions) involved in running these generative art systems and are therefore unable to make informed decisions regarding their behavior. A single person running Stable Diffusion on their home computer might not significantly impact energy consumption, but at the scale at which these systems are currently being utilized, the impact is magnified substantially. This approach also allows us to frame this problem as an issue of education and awareness, without pointing fingers. Ideally, future cross-disciplinary approaches can lead to solutions on how to raise awareness among the different stakeholders.

Next Steps for Stakeholders

In this section, we would like to briefly outline some possible next steps for different stakeholders who are involved in cAI.

Users

As of this moment, most end users of these systems appear to be "everyday" users, rather than professional users who apply these systems in larger commercial settings. While we wager that this is likely to change in the future, we still think it is worth engaging these current users in the larger discourse on sustainable practices regarding digital technologies. The issues surrounding digital overconsumption and digital waste, while evident in the current usage trends of generative AI art systems, go far beyond this single technology. If we want to foster long term sustainable tech practices among users, sustained efforts in outreach and education will be required. While it is unrealistic to expect radical and rapid change in user behavior even with targeted education campaigns, we do believe that user awareness and participation in the discourse on the environmental impact of technology has the potential to create pressure on developers and encourage the creation of more sustainable systems.

Developers

Developers of generative AI art systems, in industry and academia alike, must seriously consider the computational costs of their work, not only from a monetary perspective associated during training but also in relation to its environmental impact during inference. Discussions on Green AI have been part of the current discourse for several years at this point. Green AI, as defined by Schwartz, Dodge, Smith and Etzioni (2019), refers to improvements in AI that occur without an increase in required computational costs. While an in-depth discussion of Green AI, as well as greenwashing within the ICT (Information and Communications Technology) sector, is not within the scope of this paper, we do want to take the time to advocate for increased transparency, particularly from larger corporate entities that are now making the mass adoption of generative AI art systems possible. It is important to acknowledge that some efforts are already being made here: the environmental impact of Stable Diffusion models during training is already being estimated using the Machine Learning Impact Calculator (Lacoste et al. 2019) and made available to the public (HuggingFace n.d.). However, in order to create a clearer picture of the true environmental impact, more data is needed to accurately estimate energy consumption and carbon emissions, particularly during inference. Some recent reports have demonstrated that across the ICT sector, carbon footprint estimates have been significantly under-reporting true emission levels (Freitag, Berners-Lee, Widdicks, Knowles, Blair and Friday 2020), highlighting the need for more reliable and transparent data.

Research Community

While the focus of our paper has been on the recent rise of generative AI art systems, we are hardly the first to attempt to raise awareness on the climate implications of AI (and the ICT sector more broadly). Instead of reiterating best practices around the development of more sustainable (generative) AI systems (for examples see: Lacoste et al. 2019; Schwartz et al. 2019; Luccioni and Hernandez-Garcia 2023), we would like to commend current on-going research efforts and encourage the creation of more spaces where these important conversations can take place. There have already been a number of conferences this year alone (such as ICCC, AIES and CVPR) that offer either special tracks and/or workshops for research on (generative) AI and our climate. With the rapid current technological developments and the rise in consumer-facing generative systems it is more important than ever that the impacts of these systems, both environmental and social, are further researched and discussed.

Future Work

We cannot stress enough that a lot more work needs to be done in this area and that this position paper mainly serves as a tool to engage the larger cAI-community (researchers, industry, and users) in the discussion surrounding sustainable AI and how our tools impact the environment we live in. We would like to invite researchers to consider the following areas as potential new areas of inquiry:

Expansion of Research into Climate Impacts

First and foremost, there is a need for a more comprehensive approach to the calculation of energy consumption and GHG emissions that are associated with generative art systems.

Unfortunately, we encountered frequent problems when trying to gather data to support our thoughts in this paper. There is only a small amount of publicly available data on the use of these systems and increased transparency is required for researchers to estimate the climate implications more accurately. The following data would be required for future work: 1) daily active users for each system, 2) total daily image output for each system, 3) hardware used to process data (GPUs, cloud services), 4) location of users (as the GHG emissions associated with electricity generation varies by country and energy source (Luccioni and Hernandez-Garcia 2023)) and 5) any initiatives that might have been taken by OpenAI, Midjourney and StabilityAI to reduce climate implications.

Human Interaction with Generative AI Systems

Future work should look further into the behavior patterns and motivations of users of these generative art systems.

For instance, we hypothesize that one possible explanation for the recent explosion of use, is related to uses and gratification theory (Katz, Blumer and Gurevitch 1973), a widely cited framework to study how media has the ability to satisfy a person's needs and desires, which leads to the continued and prolonged consumption. It has already been used to explain the addictive nature of social media platforms like TikTok (Montag, Yang and Elhai 2021). Particularly, a need for escapism has been linked to increased consumption of digital content (Omar and Dequan 2020). It should be investigated whether these ideas also apply in the context of creative AI tools which allow almost instant content creation and consumption, which would not be possible otherwise. These tools allow users to create artworks at a quality that would have required high levels of fine arts skills and a significant time investment (both to acquire the skills and then to execute the artwork). Further study of how these tools interact with our cognitive reward system also demands heightened attention as these tools move from mainly still imagery to full video creation and eventually 3D VR environments. We have conjectured that there could be potential dangers here (as well as opportunities for social good), related to the notion of escapism and the ability of users to now create artificial worlds that serve as a virtual sanctuary from reality.

While this is an issue for further studies – this general hypothesis around these systems having the potential of satisfying essential human needs such as personal/creative expression and connectedness, does speak to their explosive growth and the possible issue of encountering increased difficulty and resistance in trying to educate users on the negative environmental implications of their actions.

Due to the widespread adoption of these tools, such work could also provide valuable insights into everyday human-AI interaction and meaning making in the digital space. This could further our understanding of the societal impact of these new technologies.

Limitations

We made our best efforts to find up-to-date data to back up our analysis and calculations. However, it is important to acknowledge that it is currently difficult to obtain a complete picture of the demonstrated problems due to a lack of availability of reliable data. In the Future Work section, we outlined the kinds of data that would be required to get more precise estimates on the climate implications. It is also important to discuss the possibility that the current numbers that we presented in this work merely reflect a snapshot of user behavior during a time of immense "hype" around these systems due to their novelty. However, while it is most likely correct to assume that the current level of fascination that many users have with these systems will die down eventually, these systems are likely to be adopted within many professional creative contexts in the future and while there might be a demographic shift in the users base and the types of application we encounter, we predict that the overall use of generative AI art systems will likely increase over the month and years to come.

References

Ayed, R.B. and Hanana, M. 2021. Artificial Intelligence to improve the food and agriculture sector. *Journal of Food Quality*. https://doi.org/10.1155/2021/5584754

Bietti, E. and Vatanparast, R. 2020. Data waste. *Harvard International Law Journal* 61:11 pages

Bisshop, L.; Hendlin, Y.; and Jaspers, J. 2022. Designed to break: planned obsolescence as corporate environmental crime. *Crime, Law and Social Change* 78*:*271-293

Brown, P.M. and Cameron, L.D. 2000. What can be done to reduce overconsumption? *Ecological Economics* 32(1):27-41

Burnes, A. 2022, Nov 15. A deeper look at power consumption: GeForce RTX 4080 power is about more than TGP. NIVIDIA. https://www.nvidia.com/en-my/geforce/news/geforce-rtx-40-series-ultra-efficient-beyond-fast/

Ceci, L. 2023a, Jan 27. Lensa AI app downloads worldwide [Infographic]. Statista. https://www.statista.com/statistics/1350961/lensa-ai-app-downloads-worldwide/

Ceci, L. 2023b, Jan 27. Lensa AI in-app purchase revenue 2021-2022 [Infographic]. Statista. https://www.statista.com/statistics/1350980/lensa-ai-in-app-revenue-worldwide/

Cervenka, E. 2022, Oct 5. All you need is one GPU: Inference benchmark for stable diffusion. LambdaLabs. https://lambdalabs.com/blog/inference-benchmark-stable-diffusion

Chen, A.; Dietrich, K.N.; Huo, X.; and Ho, S. 2011. Developmental neurotoxicants in e-waste: An emerging health concern. *Environmental Health Perspectives* 119(4): https://doi.org/10.1289/ehp.1002452

deVries, A. 2019. Renewable energy will not solve Bitcoin's sustainability problem. *Joule* 3:893-898

deVries, A. 2022. Cryptocurrencies on the road to sustainability: Ethereum paving the way for Bitcoin. *Patterns* 4(1): 100633

Dwivedi, Y.K.; Hughes, L.; Kar, A.K.; Baabdullah, A.M.; Grover, P.; Abbas, R.; Andreini, D.; Abumoghli, I.; Barlette, Y.; Bunker, D.; Kruse, L.C.; Constantiou, I.; Davison, R.M.; De', R.; Dubey, R.; Fenby-Taylor, H.; Gupta, B.; He, W.; Kodama, M.; Mäntymäki, M.; Metri, B.; Michael, K.; Olaisen, J.; Panteli, N.; Pekkola, S.; Nishant, R.; Raman, R.; Rana, N.P.; Rowe, F.; Sarker, S.; Scholtz, B.; Sein, M.; Shah, J.D.; Teo, T.S.H.; Tiwari, M.K.; Vendelø, M.T.; and Wade, M. 2022. Climate change and the COP26: Are digital technologies and information management part of the problem or the solution? An editorial reflection and call to action. *International Journal of Information Management* 63:102456

Energy Information Administration. International – Electricity. Eia.gov. https://www.eia.gov/international/data/world/electricity/electricity-consumption

Fatunde, M. and Tse, C. 2022, Oct 17. StabilityAI raises seed round at $1 billion value. Bloomberg. https://www.bloomberg.com/news/articles/2022-10-17/digital-media-firm-stability-ai-raises-funds-at-1-billion-value

Freitag, C.; Berners-Lee, M.; Widdicks, K.; Knowles, B.; Blair, G.; and Friday, A. 2020. The climate impact of ICT: A review of estimates, trends and regulations. arXiv:2102.02622

Hatmaker, T. 2022, Dec 5. Lensa AI, the app making 'magic avatars,' raises red flags for artists. TechCrunch.

https://techcrunch.com/2022/12/05/lensa-ai-app-store-magic-avatars-artists/

Huggingface. Stable Diffusion v1-4 Model Card. Hugging Face. https://huggingface.co/CompVis/stable-diffusion-v1-4#environmental-impact

Huntingford, C.; Jeffers, E.S.; Bonsall, M.B.; Christensen, H.M.; Lees, T.; and Yang, H. 2019. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters* 14:124007

Jobba, L. and Herweijer, C. How AI can enable a Sustainable Future. https://www.pwc.co.uk/sustainability-climate-change/assets/pdf/how-ai-can-enable-a-sustainable-future.pdf

Katz, E.; Blumler, J.G.; and Gurevitch, M. 1973. Uses and gratification theory. *Public Opinion Quarterly* 37:509-523

Kelly, K. 2022, Nov 17. Picture limitless creativity at your fingertips. Wired. https://www.wired.com/story/picture-limitless-creativity-ai-image-generators/

Ksherti, N. and Voas, J. 2022. Blockchain's carbon and environmental footprints. *Computer* 55(8):89-94

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the carbon emissions of machine learning. arXiv:1910.09700v2

Lewis, N. 2022, Dec 10. Here's the PC hardware you should buy for Stable Diffusion. HowToGeek. https://www.howtogeek.com/853529/hardware-for-stable-diffusion/

Lim, T. 2022, Nov 30. Google Play's best apps and games of 2022. Google - The Keyword. https://blog.google/products/google-play/google-plays-best-apps-and-games-of-2022/

Luccioni, A.S. and Hernandez-Garcia, A. 2023. Counting carbon: A survey of factors influencing the emissions of machine learning. arXiv:2302.08476v1

Montag, C.; Yang, H.; and Elhai, J.D. 2021. On the psychology of TikTok use: A first glimpse from empirical findings. *Frontiers in Public Health* 9: https://doi.org/10.3389/fpubh.2021.641673

Mostaque, E. [@EMostaque]. 2022, Aug 22. New Midjourney beta using Stable Diffusion looking absolutely awesome [Tweet]. Twitter. https://twitter.com/EMostaque/status/1561917541743841280?lang=en

Moutaib, M.; Fattah, M.; and Farhaoui, Y. 2020. Internet of things: Energy consumption and data storage. *Procedia Computer Science* 175:609-614

OpenAI. 2022, Nov 3. Dall-E API now available in public beta. OpenAI. https://openai.com/blog/dall-e-api-now-available-in-public-beta/

Omar, B. and Dequan, W. 2020. Watch, share or create: The influence of personality traits and user motivation on TikTok mobile video usage. International Association of Online Engineering. https://www.learntechlib.org/p/216454/

Pennington, A. 2022, Nov 29. You might actually find a muse in the machine (or Midjourney). NABAmplify. https://amplify.nabshow.com/articles/ic-the-muse-in-the-machine/

Perez, S. 2022, Dec 12. AI art apps are cluttering the app store's top charts following Lensa AI's success. TechCrunch. https://techcrunch.com/2022/12/12/ai-art-apps-are-cluttering-the-app-stores-top-charts-following-lensa-ais-success/

Perkins, D.N.; Drisse, M.B.; Nxele, T.; and Sly, P.D. 2014. E-waste: A global hazard. *Annals of Global Health* 80(4):286-295

PRNewswire. 2023, Feb 07. StabilityAI partners with KrikeyAI to launch AI animation tools. PRNewswire.com. https://www.prnewswire.com/news-releases/stability-ai-partners-with-krikey-ai-to-launch-ai-animation-tools-301739683.html?tc=eml_cleartime

Schwartz, R.; Dodge, J.; Smith, N.A.; and Etzioni, O. 2019. Green AI. arXiv:1907.10597v3

Singh, N. and Ogunseitan, O.A. 2022. Disentangling the worldwide web of e-waste and climate change co-benefits. *Circular Economy* 1:100011

StabilityAI. 2022, Oct 17. Stability AI Announces $101 Million in Funding for Open-Source Artificial Intelligence. StabilityAI. https://stability.ai/blog/stability-ai-announces-101-million-in-funding-for-open-source-artificial-intelligence

StabilityAI. [@StabilityAI]. 2023, Feb 16. We're glad to be partnering with KrikeyApp [Tweet]. https://twitter.com/StabilityAI/status/1626164348220579840?cxt=HHwW-gIC9ofz2pZEtAAAA

StabilityAI. General FAQ. StabilityAI. https://stability.ai/faq

StabilityAI. Stable Diffusion Launch Announcement. StabilityAI. https://stability.ai/blog/stable-diffusion-announcement

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. arXiv:1906.02243v1

Walton, J. 2023, Jan 25. Stable Diffusion Benchmarked: Which GPU Runs AI Fastest (Updated). TomsHardware. https://www.tomshardware.com/news/stable-diffusion-gpu-benchmarks

Widmer, R.; Oswald-Krapf, H.; Sinha-Khetriwal, D.; Schnellmann, M.; and Böni, H. 2005. Global perspectives on e-waste. *Environmental Impact Assessment Review* 25(5):436-458

Wombo.AI [@wombo.ai]. 2021, Nov 16. Happy to announce the release of WOMBO Dream! [Instagram Post]. Instagram. https://www.instagram.com/p/CWWLMUGJTBB/?hl=en

Wombo Inc. About – Imagine new worlds with AI. Wombo Inc. https://www.w.ai/about

How to Make AI Artists Feel Guilty in a Good Way? Designing Integrated Sustainability Reflection Tools (SRTs) for Visual Generative AI

Yuanyang Ren[*] and **Aswath Sivakumaran**[†] and **Emma Niemelä**[‡] and **Petra Jääskeläinen**[§]
MID, EECS, KTH Royal Institute of Technology, Stockholm, Sweden
yren@kth.se, aswaths@kth.se, emmanie@kth.se, ppja@kth.se

Abstract

AI can be energy intensive, and artists currently lack access to empowering information. With growing concerns of climate change and calls for environmental sustainability, there is a real need to explore strategies to communicate sustainability information to artists using generative AI, given its increasing presence and widening accessibility. This paper presents an exploratory Research-through-Design study (including design-informing survey, design prototyping, user testing) of integrating sustainability reflection features into generative AI systems, and provides preliminary knowledge of the design characteristics that can be leveraged, including artists' experiences of them. This paper finds that granular, relatable data visualizations and informed use of colors are effective in communicating about energy consumption. Furthermore, artists were positive towards ”feeling bad” in the process of becoming aware of their impacts, and called for systems that could provide them low-energy settings during exploratory stages of the artistic process.

Introduction

Artificial Intelligence (AI) is computationally intensive (Dhar 2020) with increasing power demands (Mehonic and Kenyon 2022). However, the complex and often black-box nature of AI makes it difficult for people to understand the environmental impact of their work. The recent rapid and increasing use of accessible Generative AI (G-AI) tools warrant investigation into the sustainability aspects of such technology. This paper explores how to design **Sustainability Reflection Tools (SRTs)** for **Visual Generative AI (VG-AI)**. This paper serves as an exploratory study into transparentizing the environmental impact of AI art generators and discusses how designers (and users) of such tools can take steps to addressing the environmental sustainability problems glooming in the horizon.

State of Research: SRTs for Generative AI

Self-reflection for sustainability has been widely researched in HCI (Kefalidou et al. 2015), with a particular focus on studies and tools that attempt to promote more pro-environmental energy consumption habits. Strategies have included; information displays through smart monitors (Froehlich, Findlater, and Landay 2010), personalized information delivery (Mankoff et al. 2007), or computer games (Bang, Torstensson, and Katzeff 2006). Furthermore, well-established methodologies such as Life Cycle Assessment (LCA) and other critical frameworks (Grover, Emmitt, and Copping 2019) have been explored to analytically make sense of sustainability of technologies.

However, despite this diversity of research and tools development in HCI, the current state of SRTs in the context of AI - and particularly G-AI - is lacking both in terms of research and practically usable tools. Existing research focuses on building tools for the more technically inclined (Anthony, Kanding, and Selvan 2020), or is more generic (Lacoste et al. 2019) rather than geared towards specific applications. Simultaneously, current research has brought up environmental sustainability concerns in the specific context of G-AI (Jääskeläinen, Pargman, and Holzapfel 2022a; Jääskeläinen, Pargman, and Holzapfel 2022b; Bender et al. 2021), although attempts at addressing these concerns practically are scarce. Currently there are no SRTs aimed specifically for G-AI, or AI artists. However, as discussed previously, current SRTs are not suitable for non-technical end users (who in this case may be professional artists, or any individuals engaging in image-making using generative tools) and not necessarily well-versed in AI, Computer Science, or Environmental studies. The combination of the increasing energy demands of AI and G-AI tools becoming more prevalent and widely accessible (regardless of technical skill or available hardware) warrants the necessity of empowering as many stakeholders along the line as possible to take control (or at least be informed) of their environmental impact created while using these tools.

SRTs and The Complexity of Behavior Change for Sustainability

Multiple models exist to promote pro-environmental behavior. In this paper we have employed the most used ”information” model (Froehlich, Findlater, and Landay 2010), and augmented it using colors, pictograms, and data visualizations. However, research surrounding what shapes pro-environmental or pro-sustainability behavior is not clear

[*] 1st author: prototype design, created & performed user tests.
[†] 2nd author: survey design, co-created user tests, co-writer.
[‡] 3rd author: contributed to user tests.
[§] 4th author: project supervisor, co-writer.

(Kollmuss and Agyeman 2002). Further, no long-term studies about the effects pertaining to SRTs or similar reflection tools exist, while short-term studies (mostly on carbon calculators) are heavily criticized, lack empirical evidence, and are mostly inconclusive (Biørn-Hansen, Barendregt, and Andersson 2020). This is primarily due to the simplistic viewing of change towards sustainability when in reality behavior change is complex (Brynjarsdottir et al. 2012; Strengers 2014) and influenced by various factors, such as the context, prior knowledge, feelings, culture, etc. However, many SRT related studies build on an underlying assumption that presenting information is enough to result in behavior change, painting a picture of humans as more consciously rational agents than they likely are. However, studies of eco-feedback tools have shown some promise (Holmes 2009) and warrant further investigation of what specific conditions and factors are successful (or not) in facilitating behavior change, in specific use practices. This exploratory empirical study *we do not argue or aim for long-term behavior change*, but rather focused on developing knowledge on these underlying factors that lay ground on behavior change for sustainability in context of VG-AI through asking: (1) How do certain design characteristics (colors, symbols, infographics) and strategies relate to the effectiveness of *communicating* sustainability-related information?, (2) What kind of quality v. impact trade-offs are users willing to make?, (3) What insights/themes can we draw from participants' experiences in the user tests to inform future research and development in VG-AI SRTs?

Methods

To address our research questions, we used an exploratory Research-through-Design (RtD) approach (Zimmerman and Forlizzi 2014) that involved a design-informing survey to map user insights relating different aspects of sustainability information representation through **Survey**, designed a **Prototype**, and performed a **User Study** in an exploratory setting.

Survey Questionnaire To Obtain General Insights of VG-AI Users

An online survey was distributed through social media and personal networks, targeted broadly at people who have used VG-AI systems. The survey consisted of 21 questions assessing attitudes and experience of color, pictographic associations, and data visualization. Additionally, demographic data (age, gender, current residence) was collected to ensure specificity and clarity of data due to the non-random convenience sampling (Gideon 2012). Background information was collected regarding respondents attitudes around environmental sustainability and familiarity around AI art generators [1], competence in art/design[2] (as our target audience was non-specific), and color blindness declaration to get an overview of factors that may influence how people experienced the evaluated aspects. Prior artistic experience was of interest, since respondents with experience in art might be more critical of color and pictograms. Eventually, data was gathered from 26 respondents, mostly located in Europe (n_1=20), followed by India (n_2=3) and New Zealand (n_3=2). Majority had some art/design competency (Advanced knowledge (student or professional)=8, Hobbyist=4, Some basic knowledge=9), 5 reported having no competency. Majority were aged 25-34 (n_4=18), the rest were between 18-24 (n_5=5) and 35-44 (n_6=2). The questions asked in the survey gathered insights specifically on these aspects (see also Fig. 1): (1) **Color associations**[3] in relation to energy consumption levels; (2) **Pictogram associations** in relation to environmental sustainability; (3) **Preferences and experience of data visualization** style for visualizing energy consumption[4].

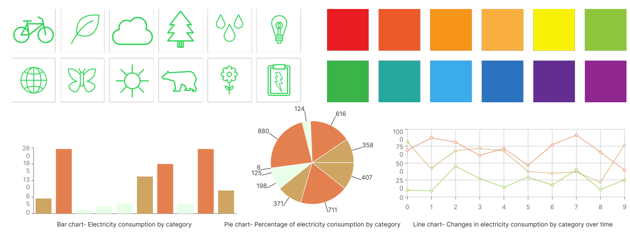


Figure 1: Colors (top-left), Pictograms (top-right), Data visualization styles (bottom).

Design Process and Prototype

A Figma prototype[5] of an AI art generator with SRT augmented features was developed (see Fig. 2). The prototype was based on Nightcafe[6] with UI modifications made with the intent to inform the user about the energy consumption of their usage. The SRT features were informed by the survey data. The prototype consists of a predetermined path for users to take during the user study.

User Study

A 3 stage user-study was conducted to understand participants' perception of AI art generators with integrated SRTs: (1) Using Nightcafe to familiarize themselves with the base UI; (2) Using our prototype while following a set of instructions and thinking aloud (Martin and Hanington 2012) to understand their thought process, emotional states, and perception; (3) Interview to evaluate specific aspects of the prototype's SRT design qualities. The interview included

[1] 5pt Likert scale: 1=Strongly Disagree (SD), 2=Disagree (D), 3=Neutral (N), 4=Agree (A) 5=Strongly Agree (SA)

[2] For simplicity, will refer to both as *art* in the rest of the paper.

[3] While color associations have been studied in the past (Elliot and Maier 2007), the lack of consistency in color relations (Adams and Osgood 1973), and lack of color study around eco-associations motivated us to include this question. Furthermore, approaching these questions from an exploratory RtD perspective, we wanted the design to rely and be informed by empirical data.

[4] Data visualizations do not contain real measurement data - we only focused on exploring visualization strategies.

[5] Prototype can be accessed at this link.

[6] Nightcafe is a well known multi-model online AI art generator.

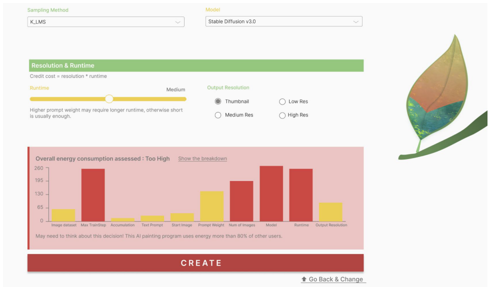


Figure 2: The SRT Prototype

questions that can broadly be divided into 3 areas of investigation: *Attention, Knowledge/Information,* and *Feeling/Emotion*. The following areas were evaluated on a Likert scale of 1(lowest) to 5 (highest): (A) Effectiveness in drawing *attention* to (UTQ_1), providing *information* about (UTQ_2) , and (UTQ_3) increasing *understanding* of the environmental impact; Anticipated effectiveness in *influencing* behavior change towards reduced consumption (UTQ_4); *Feeling* about different design elements, and emotions experienced during the whole interaction (UTQ_7); Evaluating the trade off between quality (1) or sustainability (5) (UTQ_8). 6 participants were involved in the user tests. P_1 and P_6 is from Italy, P_2 from Sweden, P_3, P_4, and P_5 from Finland. P_1 and P_3 identify as male, the rest as female. P_1, P_2, P_3, P_4, and P_6 are professional artists, while P_5 is a art hobbyist.

Results & Analysis

Survey Results

Color Associations Emphasize Visual Culture Rather than Nature 72% (n_{10}=18) reported some shade of green represented low power consumption and 60% of them (n_{11}=12) reported choosing green as they have seen it depict eco-friendliness. Only 17% (n_{12}=3) associated it with nature - e.g. *nature is green*. One respondent specifically wrote the associations were a *"cultural agreement"*. 88% (n_{15}=22) associated red with high power consumption and justified the choice by association to danger/concern/alarm (n_{17}=9), or a negative environmental labels (n_{18}=6). Similar to green, very few participants associated it with natural phenomena such as heat/warmth/fire (n_{19}=3). Respondents were split between oranges (n_{20}=10) and yellow (n_{21}=11) for middle level power consumption 47% associating orange (n_{22}=5) or yellow (n_{23}=5) with being used as a *middle color* between green and red, with some specifically recalling the orange in traffic lights (n_{24}=3). Interestingly, some picked green (n_{25}=5), including the respondent who reported being red-green colorblind; the reason for picking this color however is not entirely clear with with 2/5 responding that it was the middle/neutral color. When analyzing the results, it is evident that majority of the respondents had constructed an association through exposure to certain type of visual culture (majority reported seen it used in a similar context in society, in contrast to having seeing it in natural environments). This indicates, that color associations are dynamically constructed in contemporary culture - and rely less on natural representations.

Pictograms Rely on Symbolic Association to Nature Pictogram of a leaf was among the top 3 choices to represent sustainability due to their association with nature/environment (n_{25}=18). This is in stark opposition to color discussed in the previous section. Symbols closely associated with natural elements (flower, polar bear, sun, butterfly, drops, tree, leaf) were picked more often (37 votes for 8 pictograms) than man-made elements (energy, globe, light bulb, bicycle) (9 votes for 4 pictograms). The cloud pictogram was the only nature pictogram that was never picked, perhaps due to its greater association with *cloud computing* or *cloud storage*; there is however, no clearly discernible reason.

Preference for Granularity in Data Representations Though the bar chart was chosen by almost half the sample (n_9=12), all respondents (irrespective of graph style) noted that their choices were due to (1) ease of quick parsing, and (2) ease of understanding where and when power consumption has spiked/sunk. This aligns with literature where people want more granular transparency (Padgett et al. 2008), i.e., simply informing users about low/high consumption is not enough. Thus, future design patterns could include the provision of high granularity or customization to increase effectiveness of communicating sustainability-related information.

Insights from the User Testing

Colors and Data Visualization Are More Important than Pictograms All participants immediately understood what the colors indicated, P_6 reported the colors were all *"very universal"*. The use of red also made participants think more about their actions and made them feel bad about their high energy usage. Color was also the highest rated[7] design characteristic, both individually (Bringing Attention=4.34, Informing=4.5, Increasing Awareness=4.5, Effective in potentially changing behavior=4.5) and on average ($\mu = 4.46$); followed by data visualization (Bringing Attention=4.5, Informing=4.17, Increasing Awareness=4.17, Effective in potentially changing behavior=4.5, $\mu = 4.34$). Participants consistently reported that the data visualization *"[was] the most interesting"* (P_3), *"[made me] want to do better"* (P_2), *"gives me the curiosity to discover precisely which [choice] will change what."* (P_1). However, all participants reported wanting data to be relatable *(ref. Information Should be Relatable)*. The pictogram was barely noticed by participants and the least important (Bringing Attention=3, Informing=3.5, Increasing Awareness=3.5, Effective in potentially changing behavior=2.67, $\mu = 3.17$), with only P_4 ($\mu_{symbols} = 4.5$) and P_5 ($\mu_{symbols} = 4.75$) rating it high. P_1 said he *"was not really impressed"* and *"did not give attention to that"*, P_2 said about the leaf: *"didn't notice the...flower withering"*, P_6 *"forgot about them"*, and P_4 had

[7]All scores are out of 5. Those with recurring decimals are rounded up.

to be shown the pictogram again when asked to rate it as they forgot it completely.

Information Should be Relatable As discussed in *Colors and Data Visualization Are More Important than Pictograms*, participants were highly interested in seeing the data of their usage and choices. However, participants expressed dismay and confusion about the values provided: *"500kW....I don't know if it's a lot..."* (P_2), *"they [kW values] can be really abstracted. Okay. 100 kilowatts. So what, what does it mean?* (P_3), *"I would like to have some data ... [like] keeping your television on for two hours.."* (P_1), *"how is it in relation to all the energy consumption ... if I make a video to YouTube..."* (P_4). Providing data in relatable terms similar to Huang, O'Neill, and Tabuchi would allow for the data to be more effective in informing participants about their consumption. Participants also provided other ideas for more effective communication, such as implementing a virtual *bot* that informs you of the effects of your choices and corresponding parameters (P_3), or using emoticons (P_6).

Evoking "Negative" Emotions for Good All participants reported feeling bad about having high energy consumption being reported in the prototype. However, they all reported that this *bad feeling* was positive as it made them want to go back and change their outputs to be more *green*. P_2 compared the experience of getting red to *loosing a game*. P_1 provided the analogy of someone asking him to pickup litter from the ground: even if he feels bad about it, its for a good reason.

Systems Could Provide Low-Energy Settings for Experimentation and Exploration Phases of Creative Endeavors All participants (except P_3) wanted to go back and change their outputs to lower the energy use, though P_3 had expressed that he was limited by a static prototype. Although participants remained divided about preferring quality or sustainability ($\mu = 2.5$), they all suggested they would use lower consuming settings during initial experimentation or drafts, and use the highest quality settings for final deliverables. This insight also indicates that future work on VG-AI SRTs could take into consideration the various stages of creative work (Jääskeläinen, Pargman, and Holzapfel 2022a) and tailor capabilities and offerings to those specific stages. However, artists expressed concern over this method of working due to the randomness of VG-AI generation, and the experimental nature of creative work, *viz.* the randomness of digital noise inherent to G-AI is at odds with other more "controllable tools", such as Photoshop. However, optimistically all participants reported that they would want to consume less energy after using our prototype and became more informed about their power consumption, which shows promise regarding future work in SRTs.

Discussion & Conclusion

In this paper, we have provided preliminary knowledge on designing SRTs for VG-AI through an empirical design research study that included surveys, prototyping of an SRT tool, and user tests. To summarize the results, colors and data representations were more important than symbols in communicating sustainability information, and there was a clear connection of color association to contemporary visual culture. Pictograms, in contrast, relied on symbolic association to nature - and participants referred to ones that had a stronger association to nature. When it came to data representation, users preferred granular and relatable representation, such as comparing the consumption to *X* hours of watching YouTube videos. Participants also experienced feeling "bad", but described these emotions as something positive - and were willing to experience them in a process of becoming more aware of the sustainability implications of their work. Our study showed the artistic community responding positively to addressing environmental sustainability aspects of their practice - and that will be important going forward with the research agenda concerning SRTs. Further, these positive attitudes may help people to be, for example, more receptive to presented information, or more willing to change their behaviors; it may even be a predicament that needs to be met for SRTs designs to have net positive effects on the users. To acknowledge some of the limitations of this study, we would like to return to the complex nature of behavior change briefly discussed in *SRTs and The Complexity of Behavior Change for Sustainability*. Firstly, it is challenging to confirm if SRTs would in fact inflict behavior change without longitudinal studies with ethnographic and contextual observations of artists' work practices. Furthermore, we can anticipate that there are several factors outside the scope of SRTs that incentivize and drive people towards more or less sustainable ways of using VG-AIs. For example, it has been pointed out that higher socio-economic position likely enables people to use these systems to a greater extent, as they are primarily emerging in the Global North (Jääskeläinen, Holzapfel, and Åsberg 2022). Furthermore, one of the important tensions that should be acknowledged going forward with SRTs is that the underlying assumption that humans are rational agents and will change their behavior when prompted with information is weakly grounded. Thus, we argue that more research should be directed towards emotional aspects of SRTs and how we might inflict emotional experiences that commit users to certain practices on a deeply personal level - including the question: ***How to make artists feel guilty in a good way?***

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation (Grant 2020.0102).

References

[Adams and Osgood 1973] Adams, F. M., and Osgood, C. E. 1973. A Cross-Cultural Study of the Affective Meanings of Color. *Journal of Cross-Cultural Psychology* 4(2):135–156.

[Anthony, Kanding, and Selvan 2020] Anthony, L. F. W.; Kanding, B.; and Selvan, R. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. arXiv:2007.03051 [cs, eess, stat].

[Bang, Torstensson, and Katzeff 2006] Bang, M.; Torstensson, C.; and Katzeff, C. 2006. The PowerHhouse: A Persuasive Computer Game Designed to Raise Awareness of Domestic Energy Consumption. In IJsselsteijn, W. A.; de Kort, Y. A. W.; Midden, C.; Eggen, B.; and van den Hoven, E., eds., *Persuasive Technology*, Lecture Notes in Computer Science, 123–132. Berlin, Heidelberg: Springer.

[Bender et al. 2021] Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. Virtual Event Canada: ACM.

[Biørn-Hansen, Barendregt, and Andersson 2020] Biørn-Hansen, A.; Barendregt, W.; and Andersson, D. 2020. Introducing Financial Data and Groups in a Carbon Calculator: Issues With Trust and Opportunities for Social Interaction. In *Proceedings of the 7th International Conference on ICT for Sustainability*, 11–17. Bristol United Kingdom: ACM.

[Brynjarsdottir et al. 2012] Brynjarsdottir, H.; Håkansson, M.; Pierce, J.; Baumer, E.; DiSalvo, C.; and Sengers, P. 2012. Sustainably unpersuaded: how persuasion narrows our vision of sustainability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 947–956. New York, NY, USA: Association for Computing Machinery.

[Dhar 2020] Dhar, P. 2020. The carbon impact of artificial intelligence. *Nature Machine Intelligence* 2(8):423–425. Number: 8 Publisher: Nature Publishing Group.

[Elliot and Maier 2007] Elliot, A. J., and Maier, M. A. 2007. Color and Psychological Functioning. *Current Directions in Psychological Science* 16(5):250–254. Publisher: SAGE Publications Inc.

[Froehlich, Findlater, and Landay 2010] Froehlich, J.; Findlater, L.; and Landay, J. 2010. The design of eco-feedback technology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999–2008. Atlanta Georgia USA: ACM.

[Gideon 2012] Gideon, L., ed. 2012. *Handbook of Survey Methodology for the Social Sciences*. New York, NY: Springer New York.

[Grover, Emmitt, and Copping 2019] Grover, R.; Emmitt, S.; and Copping, A. 2019. Reflecting on sustainability: coproducing a critical framework for sustainable design in the architectural studio. *Higher Education Pedagogies* 4(1):41–63. Publisher: Routledge _eprint: https://doi.org/10.1080/23752696.2019.1624589.

[Holmes 2009] Holmes, T. 2009. *Eco-visualisation: Combining art and technology to reduce energy consumption*. Thesis, University of Plymouth. Accepted: 2013-11-20T14:22:02Z.

[Huang, O'Neill, and Tabuchi 2021] Huang, J.; O'Neill, C.; and Tabuchi, H. 2021. Bitcoin Uses More Electricity Than Many Countries. How Is That Possible? *The New York Times*.

[Jääskeläinen, Holzapfel, and Åsberg 2022] Jääskeläinen, P.; Holzapfel, A.; and Åsberg, C. 2022. Exploring More-than-Human Caring in Creative-Ai Interactions. In *Nordic Human-Computer Interaction Conference*, 1–7. Aarhus Denmark: ACM.

[Jääskeläinen, Pargman, and Holzapfel 2022a] Jääskeläinen, P.; Pargman, D.; and Holzapfel, A. 2022a. On the Environmental Sustainability of Ai Art(s). In *Eighth Workshop on Computing within Limits 2022*. LIMITS.

[Jääskeläinen, Pargman, and Holzapfel 2022b] Jääskeläinen, P.; Pargman, D.; and Holzapfel, A. 2022b. Towards sustainability assessment of artificial intelligence in artistic practices. arXiv:2210.08981 [cs].

[Kefalidou et al. 2015] Kefalidou, G.; Skatova, A.; Shipp, V.; and Bedwell, B. 2015. The Role of Self-Reflection in Sustainability. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '15, 1030–1033. New York, NY, USA: Association for Computing Machinery.

[Kollmuss and Agyeman 2002] Kollmuss, A., and Agyeman, J. 2002. Mind the Gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research* 8(3):239–260. Publisher: Routledge _eprint: https://doi.org/10.1080/13504620220145401.

[Lacoste et al. 2019] Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the Carbon Emissions of Machine Learning. arXiv:1910.09700 [cs].

[Mankoff et al. 2007] Mankoff, J.; Matthews, D.; Fussell, S. R.; and Johnson, M. 2007. Leveraging Social Networks To Motivate Individuals to Reduce their Ecological Footprints. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, 87–87. ISSN: 1530-1605.

[Martin and Hanington 2012] Martin, B., and Hanington, B. M. 2012. *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Beverly, MA: Rockport Publishers, digital ed edition.

[Mehonic and Kenyon 2022] Mehonic, A., and Kenyon, A. J. 2022. Brain-inspired computing needs a master plan. *Nature* 604(7905):255–260. Number: 7905 Publisher: Nature Publishing Group.

[Padgett et al. 2008] Padgett, J. P.; Steinemann, A. C.; Clarke, J. H.; and Vandenbergh, M. P. 2008. A comparison of carbon calculators. *Environmental Impact Assessment Review* 28(2):106–115.

[Strengers 2014] Strengers, Y. 2014. Smart energy in everyday life: are you designing for resource man? *Interactions* 21(4):24–31.

[Zimmerman and Forlizzi 2014] Zimmerman, J., and Forlizzi, J. 2014. Research Through Design in HCI. In Olson, J. S., and Kellogg, W. A., eds., *Ways of Knowing in HCI*. New York, NY: Springer. 167–189.

Shattering Bias: A Path to Bridging the Gender Divide with Creative Machines

Juliana Shihadeh and Margareta Ackerman
Department of Computer Science and Engineering
Santa Clara University, Santa Clara, California
{jshihadeh,mackerman}@scu.edu

Abstract

The widespread emergence of creative machines poses a significant challenge, as they tend to reinforce biases, including gender bias. This paper presents a novel perspective on how creative machines can be utilized to counteract gender disparities and mitigate bias. We propose research directions that explore the potential of creative systems to empower women and promote gender equity. Our aim is to leverage computational creativity to actively contribute to fostering a more inclusive and equitable society.

Introduction

The emergence of generative AI in the world at large came with a high price: The most widespread generative technologies are riddled with bias. Multiple biases have been identified in large generative models including biases related to culture (Saharia et al. 2022) and religion (Abid, Farooqi, and Zou 2021). Gender bias, which is our focus here, has been found in such models with respect to occupations (Cheong et al. 2023), intellect (Shihadeh et al. 2022), and leadership (Lucy and Bamman 2021). Millions of people are using creative machines of the likes of Dall-E (Q.ai 2023), Midjourney (Salkowitz 2022) and ChatGPT (Milmo 2023), which perpetuate and amplify these biases. In turn, the dissemination of content created with these systems further amplifies the biases inherent in these models.

Is is well known that nearly half of the world population consists of women.[1] Gender biases have a profound impact on the potential of nearly half of the global population and limit the contributions women can make compared to if they were granted equal voice, say, and opportunity. Furthermore, eradicating gender bias in large models is proving challenging as seen in the sparsity of research on solutions to this issue, challenges in accessing datasets and limited computation power to retrain models (Berg et al. 2022). Doubtlessly, research into how to reduce bias in large language models (Liu et al. 2022) and text-to-image generators (Orgad, Kawar, and Belinkov 2023) despite these challenges is critical to mitigating the societal damage that these models are inflicting on our world. In this paper, however, we offer a complementary approach.

What if, instead of perpetuating and amplifying biases, creative machines could instead be utilized to bring about a more just world? In this paper, we propose a computational creativity research agenda to promote gender equity. Grounded in interdisciplinary research spanning psychology, arts, computer science and gender studies, we offer a roadmap for a research program that has the potential to both empower women and bring profound awareness to gender bias.

The CC community carries unparalleled expertise in creative machines. In a world where creative machines are rapidly integrated into the very fabric of society, the risk of these machines to amplify biases is substantial, if not inevitable. Our community possesses the necessary expertise to envision and pioneer novel approaches for integrating creative machines into the world at large. We hope that the ideas in this paper encourage the CC community to delve deeper into the use of creative machines for fostering gender equality.

We begin this paper by briefly discussing the presence of gender bias in creative machines and the impact of gender bias on society at large. Next, we propose future directions of research in CC to mitigate gender bias along with guidelines for maneuvering the development of creative machines that close, rather than widen, the gender gap.

Gender Bias in Creative Machines

Due to their large (and as such difficult to prune) data sets, large language models and text-to-image generators are highly problematic when it comes to bias. For instance, one study found that when Stable Diffusion was prompted with "a photo of the face of" an emotional or exotic person, more women were seen (Bianchi et al. 2022). Another study found when writing stories with GPT-3, more women were associated to family, appearance and less leadership (Lucy and Bamman 2021). Studies on brilliance bias found that GPT-3 affiliated higher intellectual abilities to men (Shihadeh et al. 2022) and associated brilliance to images of men more often than women (Shihadeh and Ackerman 2023). Multiple studies look at gender stereotypes in professions, whereby the models assume that, for example, nurses must be female and doctors must be male (Cheong et al. 2023; Bianchi et al. 2022; Kirk 2021; Caliskan, Bryson, and Narayanan 2017). In the Computational Creativity commu-

[1] https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

nity, there has not been much focus on discriminatory bias, with the exception of Loughran's discussion of both algorithmic and discriminatory bias in AI (Loughran 2022).

The Effects of Biases

Biases cause *stereotype threat*, feelings of anxiety and stress that result from not feeling that one belongs (Calaza 2021). Stereotype threat further causes social isolation, rejection, and reduced memory which negatively impact one's well-being, immune system, and work quality. In a related study (Calaza 2021), female participants were given one of two math tests, one stating that their performance was to study gender differences and another focusing on cognitive processes. Women's resulting scores were lower in the former setting. Additionally, biases induce stereotypical occupational fit (Kirk 2021). When repeated through multiple sources, biases further influence the world and increase inequalities (Calaza 2021). Based on cultivation theory, repeated content exposure amplifies this effect (Potter 1993).

Research Directions

In this section, we propose avenues for computational creativity research to contribute towards creating a more gender-inclusive world. Our objective is for 1) young women to explore their potential rather than succumb to society's biases on this matter, and 2) help others gain deeper insight into women's experiences around gender bias.

Helping women realize their potential

Using creative machines, we can demonstrate that the "glass ceiling" is breakable and give girls the space to explore their true potential despite societal biases that often limit their ambitions. As Melanie Perkins, CEO of Canva - a suite of design tools - emphasizes, it is important to dream, to dream ten years out even and dream of what kind of world you want to live in (Ventures 2022).

It is important to help girls and women "close their imagination gap" [2] and show them that there are no limits to what they can pursue. For example, we can start by helping girls explore career paths. It has been identified that images "put ideas into your head" (Hibbing and Rankin-Erickson 2003). Minorities are more negatively affected by media stereotypes (Appel and Weber 2021) and girls have been found to replicate stereotypical behaviour (Essig 2018). If visuals influence people's view of the world, we could instead offer up creative visuals that allow girls to consider a more inclusive world. To this end, a co-creative system can visualize a young girl as a grown women across a wide range of professions, offering stories to flesh out the kind of impact that she may have.

Role model intervention works by having a positive representation of a minority to demonstrate that the stereotype doesn't always hold (Eschenbach 2014). Role models can have a positive impact even when their influence is conveyed through text rather than in-person interaction (Eschenbach 2014). Machines that create role models could make it more

[2]https://www.careergirls.org/about/

(a) (b) (c) (d)

Figure 1: These images, made with Midjourney, demonstrate the power of visuals for engaging our imagination in "what if" scenarios that can support gender equity and build empathy towards the challenges faced by women. (a) visualizes what breaking the "Glass Ceiling" could look like, in a world where even God is viewed as a woman, whereas (b), (c), (d) showcase female entrepreneurs, computer scientists and politicians, respectively. Creative machines can help us step into a world where such representation were the norm, which can open up the imagination of women towards their own futures and foster empathy through demonstrating what it is like to be a minority.

accessible for girls to see, hear, or even engage with role-models. As a result, this can reach a larger audience of girls and have a bigger impact on girls' trajectory. Evidence shows that female role models in counter stereotypical roles encourage younger female students to pursue such careers including politics, science, and engineering (Olsson 2018). A CC system can be designed to help young girls create their own personalized role models, combining images and text. The role model might take on some of the values, interests, or personality traits of the girl. The added level of engagement, perhaps through a chatbot representing the generated role model, may contribute to the value of such an experience.

Additionally, other individuals, such as writers, could use a role model creator to influence and inspire their stories. Similarly, systems that help create images that promote a gender equitable world can be incorporated in children's books, as their images have been found to influence female stereotypes (Hamilton et al. 2006).

Helping women cope with gender discrimination

Offering various therapeutic means to process one's experience with gender discrimination that they have experienced can help women reduce its impact. In collaboration between computational creative practitioners and therapists, computational creativity demonstrates potential to be a means for having positive psychological influence (Pease et al. 2022). In particular, Pease et al. (Pease et al. 2022) discuss how the arts have helped soldiers overcome war injuries, enable people to reconnect with their true self, restore their sense of expertise and self-esteem, and help people create a "tangible expression of who [they] are and what [they] do." Women can work with co-creative machines to write songs, poetry and make art about their challenges with sexism. This can even be more beneficial in a co-creative context that involves creative machines and multiple people (Pease et al. 2022).

Earlier generations saw many women with limited choices, often forced to remain at home as caretakers instead of pursuing their intellectual and personal aspirations beyond the boundaries of the home. Creative machines could enable women to explore alternative life paths, considering what their lives could have been like had they not been limited by gender stereotypes. In doing so we can capture the lost voices of what happens when women are not empowered to reach their true potential. This can provide a self-reflective opportunity to come to acceptance with what was lost and help with expressing and validating their identifies (Pease et al. 2022).

Lastly, we can imagine machines designed to collect and creatively, anonymously share the experiences of women. Women could tell their stories related to struggles with gender bias, and the machine could create representative, anonymous stories that capture common experiences shared by women. The resulting stories may be shared not only through text, but through any relevant modality, such as film. Both sharing of experiences and learning of others' struggles has the potential to offer therapeutic value.

Fostering Empathy

Empathy is the ability to put oneself in another's place, to see and feel the world as they do (Rusu and others 2017). The arts can help facilitate experiences of empathy (Rusu and others 2017; Bollmer 2017; Pozo 2018). By offering immersive, multi-modal, and interactive experiences that envision a world where women outnumber men in domains where biases persist, such as leadership positions and STEM, we can enable people of all genders to empathize with the biases women encounter. To create a more personalized experience, the systems could consider the user's point of view, such as their profession and hobbies, in the creation of the experience.

For example, a user may be presented with various generative, personalized scenes and prompted by the machine with follow up questions to reflect on what they saw. The generated scenarios may comprise, for instance, the opportunity to participate in a meeting where the majority of engineers are women, a congressional session where most of the politicians are female, or a meeting with a leading venture investment firm that primarily supports female founders.

Additionally, we could develop creative machines that help demonstrate what it would be like to speak in female rather than male. For example, what if language defaults to female pronouns, causing men to experience gender exclusion in pronouns? Through linguistic-centered creative experiences, we can help portray injustices that women regularly experience in the world. For instance, consider the statement "girls are as good as boys at math", which subtly expresses the idea that boys are typically better in math. What if this statement became "boys are as good as girls at math"? Neil Armstrong's quote "That's one small step for man, one giant leap for mankind." On the contrary it could be, "That's one small step for a woman, one giant leap for womankind." For example, a CC system could be a collaborative partner to create a gender-flipped dialogue that can then be brought to life as a movie scene created by the machine. Similarly, we can imagine a co-creative system for the creation of songs or stories where stereotypes are inverted. Stepping into this type of alternative reality can foster profound empathy for the realities that women face on a daily basis, and can help people of all genders become allies in the struggle towards closing the gender gap.

Guidelines

We propose several guidelines to consider when tackling gender bias through a computational creativity lens.

1. **Evoke personal connection and reflection.** In designing creative machines for closing the gender gap, it is worth considering how the machine might elicit personal connection and reflection. For instance, working with co-creative machines can facilitate introspection on personal experiences, interests, goals, or aspirations. Similarly, an immersive experience may be designed to encourage us to notice instances of gender bias in the future that we may have otherwise missed.

2. **Facilitate creative engagement**. Consider designing experiences where users are called to tap into their own creative capabilities as they explore what is possible (for themselves or others) outside of stereotypical gender roles. By enabling the user to be creative, they become more personally invested and engaged in the experience, as active engagement has been proven to enhance brain development and learning (Immordino-Yang, Darling-Hammond, and Krone 2019).

3. **Constructive and non-judgemental**. We recommend that creative systems that aim to counter gender bias be designed in an inherently constructive and non-judgemental manner, as to put the user at ease and reduce defensiveness. Living in a world where gender bias is ubiquitous, all of us carry implicit gender biases. Recognizing bias within oneself can be challenging and initially upsetting, be it biases against oneself (ex. stereotypical beliefs that limit one's own potential) or others. As such, systems that aim to correct gender biases should be designed in a manner that is respectful to the sensitive nature of this issue.

Conclusions

In this paper, we outline research directions that invite the Computational Creativity community to take the lead in steering creative machines away from widening the gender gap, and instead utilizing them towards reducing gender biases. We present the opportunity to create CC systems that lead users into personalized, creative, and immersive experiences that showcase a vision of the world we aspire to, and invite people of all genders to experience the current realities of living as a woman. In order to promote gender equity, creative machines can assist in widening women's insight into their potential while also aiding others in relating to gender disparities that women have endured in the past and continue to face presently. By embracing the potential of computational creativity to address gender biases, we can bring forth creative machines as a powerful means to break down barriers, promote gender equity, and advance towards a more diverse and equitable future.

While this work focuses on addressing bias against women, we hope that the ideas here may spark broader interest on how creative machines may be utilized to reduce discriminatory biases in a broader sense. In particular, the research directions we propose may be adapted across the gender spectrum. Further, while our focus here is on bias stemming from gender, our proposed research directions may inspire ideas for how to utilize creative machines to reduce bias based on race, age, sexual orientation, etc. Our aspiration is that this work stimulates research on the utilization of CC in reducing discriminatory prejudice, not only aiming to minimize bias in current creative machines but also employing new and imaginative methods to build creative machines that are purposefully designed to contribute to a more fair and just society.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Appel, M., and Weber, S. 2021. Do mass mediated stereotypes harm members of negatively stereotyped groups? a meta-analytical review on media-generated stereotype threat and stereotype lift. *Communication Research* 48(2):151–179.

Berg, H.; Hall, S. M.; Bhalgat, Y.; Yang, W.; Kirk, H. R.; Shtedritski, A.; and Bain, M. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*.

Bollmer, G. 2017. Empathy machines. *Media International Australia* 165(1):63–76.

Calaza, Karin C & Erthal, F. . P. M. G. . M. K. C. . D. V. T. . D. I. . C. H. C. . V. M. D. . M. L. B. . S. J. B. . o. 2021. Facing racism & sexism in science by fighting against social implicit bias: A latina & black woman's perspective. *Frontiers in Psychology*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–186.

Cheong, M.; Abedin, E.; Ferreira, M.; Reimann, R. W.; Chalson, S.; Robinson, P.; Byrne, J.; Ruppanner, L.; Alfano, M.; and Klein, C. 2023. Investigating gender and racial biases in dall-e mini images.

Eschenbach, Elizabeth A & Virnoche, M. . C. E. M. . L. S. M. . C. M. M. 2014. Proven practices that can reduce stereotype threat in engineering education: A literature review. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*. IEEE.

Essig, L. W. 2018. *A Content-Analytic Meta-Analysis of Gender Stereotyping in Screen Media*. Brigham Young University.

Hamilton, M. C.; Anderson, D.; Broaddus, M.; and Young, K. 2006. Gender stereotyping and under-representation of female characters in 200 popular children's picture books: A twenty-first century update. *Sex roles* 55:757–765.

Hibbing, A. N., and Rankin-Erickson, J. L. 2003. A picture is worth a thousand words: Using visual images to improve comprehension for middle school struggling readers. *The reading teacher* 56(8).

Immordino-Yang, M. H.; Darling-Hammond, L.; and Krone, C. R. 2019. Nurturing nature: How brain development is inherently social and emotional, and what this means for education. *Educational Psychologist* 54:185 – 204.

Kirk, Hannah Rose & Volpin, F. . I. H. . B. E. . D. F. . S. A. . A. Y. . o. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in Neural Information Processing Systems*.

Liu, R.; Jia, C.; Wei, J.; Xu, G.; and Vosoughi, S. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence* 304:103654.

Loughran, R. 2022. Bias and creativity.

Lucy, L., and Bamman, D. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, 48–55.

Milmo, D. 2023. Chatgpt reaches 100 million users two months after launch.

Olsson, Maria & Martiny, S. E. 2018. Does exposure to counterstereotypical role models influence girls' & women's gender stereotypes & career choices? a review of social psychological research. *Frontiers in psychology*.

Orgad, H.; Kawar, B.; and Belinkov, Y. 2023. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*.

Pease, A.; Ackerman, M.; Pease, N.; and McFadden, B. 2022. A roadmap for therapeutic computational creativity. In *13th International Conference on Computational Creativity*.

Potter, W. J. 1993. Cultivation theory and research. *Human Communication Research* 19(4):564–601.

Pozo, T. 2018. Queer games after empathy: Feminism and haptic game design aesthetics from consent to cuteness to the radically soft. *Game Studies* 18(3).

Q.ai. 2023. Dall·e mini and the future of artificial intelligence art.

Rusu, M., et al. 2017. Empathy and communication through art. *Review of artistic education* (13+ 14):139–146.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35:36479–36494.

Salkowitz, R. 2022. Midjourney founder david holz on the impact of ai on art, imagination and the creative economy.

Shihadeh, J., and Ackerman, M. 2023. What does genius look like? an analysis brilliance bias in text-to-imagine models.

Shihadeh, J.; Ackerman, M.; Troske, A.; Lawson, N.; and Gonzalez, E. 2022. Brilliance bias in gpt-3. In *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, 62–69. IEEE.

Ventures, B. 2022. Sunrise australia day 1 - melanie perkins keynote. https://youtu.be/FXtGE811GCw.

Diversity and Representation in ICCC: A review of Computational Creativity publication content and authorship trends 2017-2022

Marzia Bilwani[1] and Anna Jordanous[2]
[1] Independent Researcher, Pakistan. marziabilwani@gmail.com
[2] School of Computing, University of Kent, Canterbury, Kent, UK. a.k.jordanous@kent.ac.uk

Abstract

This paper examines issues of diversity and representation in the International Conference on Computational Creativity (ICCC), evaluating the diversity of authors, content and existing creative systems. We consider: the diversity of cultural context of published systems; diversity of authors in terms of gender and culture; how the pandemic affected diversity; and trends in content of accepted papers. The study covers a period of six consecutive years around the COVID-19 pandemic (2017-22), to better understand the impact of the pandemic on diversity. The research team includes individuals from different career stages and research interests, who bring diverse perspectives to the analysis. We evidence the need for greater diversity in both authorship and content, as well as in the creative systems discussed in the proceedings. The paper concludes with recommendations for increasing diversity and representation in the field of computational creativity.

Introduction

As the world becomes more globalized and interconnected, the importance of maintaining diversity increases. Thus, in the last two decades, there has been significant discourse on ensuring diversity and representation. For example, in 2019, the percentage of top-grossing films having female protagonists more than doubled from 16 percent in 2002 to 40 percent in 2019 (Lauzen 2020). However, diversity and inclusion issues abound in academic conferences (Walters 2018).

In this paper we focus on points specific to the International Conference of Computational Creativity (ICCC) in the past six years. We consider: the diversity of cultural context of published systems; diversity of authors in terms of gender and culture; how the pandemic affected diversity; and trends in content of accepted papers. Our time period of six consecutive years since 2017 helps us consider impact of the COVID-19 pandemic on diversity and representation during the years around COVID-19 worldwide lockdowns.

This paper's team of computational creativity researchers includes people with ICCC experience ranging from none (prior to 2023) to several years' engagement with ICCC; the team spans different continents and different career points.

Importance of Diversity There are many benefits of fostering diversity. For example, in the workplace, cultural diversity has been found to lead to process gains by enhancing satisfaction and creativity (Stahl et al. 2010). The Association of Computational Creativity Task Force conducted a valuable recent study of diversity in terms of demographics of people involved in ICCC (Cunha et al. 2020). Loughran (2022) discussed biases arising due to data: systems learn and create from the data they have been trained on. She also reflected upon detrimental biases relating to demographics, particularly in the representation of women in ICCC.

The importance of diversity in the creative fields stems largely from the ability of art to preserve culture and allow individuals a form of self-expression. While the goals of computational creativity may be different from those of art, by creating systems that can enhance human creativity or are capable of being creative, it can be implied that they are indirectly fulfilling the same purpose.

Diversity becomes vital, as AI systems may tend to contribute to lesser diversity in the 'collective experiments of life' and lead to greater standardization in decision-making (Loi, Vigano, and van der Plas 2020). Computational Creativity (CC) essentially challenges this standardisation, for a unique way to curb negative societal effects of AI systems.

Analysis of existing systems

In the last six years, a total of 344 papers were published. To analyse creative systems discussed in these papers, categorisation was done by reviewing each paper manually. Categorisations in conference proceedings have changed over the years. So here we focused on classifying those papers reporting artefacts being produced or systems aiding creative production, as: Cuisine, Music, Text, Visual, Coding, Sound, Problem-solving, Dance, Theatre, Games or Other.

When analysing the systems, the top three categories that emerged were producing text-based artefacts, followed by visuals and then music. This was a trend seen across all six years with 2017 and 2021 being the exceptions. The high number of systems producing text-based artefacts could be attributed to how this category included systems generating stories, headlines, poetry, song lyrics and jokes.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Year	Num papers	Num authors	Mean authors per paper (3sf)	Num countries (continents)
2017	*34*	93	**2.74**	14 (4)
2018	46	*91*	*1.98*	16 (*3*)
2019	59	140	2.37	22 (**5**)
2020	**85**	**199**	2.34	24 (4)
2021	65	171	2.63	**30** (**5**)
2022	55	148	2.69	*16* (4)

Table 1: Numbers of papers in proceedings per year for ICCC'17-22, and corresponding number of authors involved and number of countries (and continents) represented. Highest values are in bold font, lowest in italics.

Text Within this category, for most cases, the artefacts being produced were found to be in English. A few exceptions existed with systems generating French poetry (Hämäläinen, Alnajjar, and Poibeau 2022) and Portuguese headlines (Mendes and Oliveira 2020). However, the number of these systems is quite a small proportion of the overall number of papers in this category (10 out 75 papers produce artefacts in languages other than English). However, all of these languages have European origins. One system studies Japanese popular entertainment narratives (Murai et al. 2022), but with generation of plot analyses rather than text.

Visual Arts For systems producing visual artefacts, it was observed that where systems were producing paintings it was mostly the generation of Western styles using the WikiArt dataset. Japanese-influenced art is the notable exception, with ICCC publications including a Japanese facial art database (Tian et al. 2020), Ukiyo-e stylistic generation (Tian et al. 2021) and Manga graphic novel generation (Melistas et al. 2021). Availability of datasets plays a significant part in affecting the diversity of art produced using AI algorithms (Loughran 2022; Burgdorf et al. 2022). A search for datasets using the term 'art' on Kaggle and Google (Hillier 2022) showed that the top ten results consist of artworks produced by western artists. WikiArt website contains art from 106 countries worldwide, but has 89 countries not represented (wikiart.com 2023).

Music Genres across systems producing musical artefacts tended to be limited to a few categories such as classic, jazz and pop. American and British songs, or just English songs were a popular choice (Harris, Harris, and Bodily 2020; Gordon et al. 2022). Out of 28 systems generating music, only three were found to produce Italian, Spanish and German music (Ackerman, Morgan, and Cassion 2018; Navarro and Oliveira 2018; Banar and Colton). One limitation of this evaluation, however, was that not all papers provided information on the dataset being used. Some systems were also producing instrumental music and it becomes difficult to evaluate how culturally diverse they are.

(Preliminary) gender analysis

The gender of the authors submitting a paper is also an important aspect influencing how diverse or creative systems are. Therefore, gender data was analysed for 149 people who have published in more than one year in the period under study. This excludes 467 people; the decision to take this approach was based on the reliability of the data available.

Gender data was collected from either personal knowledge or from the most probable gender being estimated from an internet search. This highlights a flaw of this work, with an assumption of gender being binary and able to be detected from how the person presents in person or their internet presence. This approach was taken as we had insufficient information to be able to determine if a person's gender was neither male nor female, if their gender is different to how they present, or if their gender has changed or is fluid. We acknowledge that this weakens our gender analysis.

In our analysis, we found that only 21 percent of writers were females, with the remainder being males. A similar ratio was also observed in a report by the computational creativity task force, in which less than one-third of Senior Program Committee (PC) and PC members were female (Cunha et al. 2020). Possible reasons could be that it is more difficult for women and non-binary/gender-fluid people to advance in their careers, perhaps due to external barriers and fewer mentors/role models of the same gender. We note (binary) gender diversity of ICCC keynote speakers.

The impact of such an imbalance would mean that datasets, algorithms and creative systems will pick up on gender biases and perpetuate those further. Even with growing awareness of gender imbalances in STEM, this is a glaring reminder of the need to consider this situation.

Geographical and cultural considerations

There can be cultural implications from the timing of deadlines and other important dates in calls for paper. Table 2 gives important dates from calls for papers, in terms of requirements from authors (submissions of review / camera-ready versions of papers). Some of these dates coincide with significant cultural dates, including:[1]

- Chinese New Year (in 2023, Jan 22)
- Lent / Easter(in 2023, February 22 - April 9)
- Holi (in 2023, March 8)
- Ramadan (in 2023, March 23 to April 20)
- Passover (in 2023, April 5 to April 13)
- Cinco de Mayo (May 5)
- Buddha's Birthday (May 26)
- Juneteenth (June 19)
- The Hajj (in 2023, June 26 to July 1)
- Rosh Hashanah (in 2023, Sept 15 - Sept 17)
- Autumnal Equinox (in 2023, Sept 23)

Deadlines are important for conference organisation. However we can recognise that some people may find it more difficult to submit to, or attend ICCC, based on the above date-based observations. While conference attendees

[1]We note that many of these dates change from year to year, and give 2023 dates as a guide, sourced from `https://www.diversityresources.com/`.

Year	Main deadlines	conference dates
2017	March 3 / May 5	June 19-23
2018	March 2 / May 12 / May 25	June 25-29
2019	Feb 28 / April 19 / May 3 / May 6 / May 17	June 17-21
2020	March 8 / May 25	~~June 29 – July 3~~ Sept 7–11
2021	April 2 / June 21 / Aug 19	Sept 14-18
2022	Feb 18 / Feb 25 / April 22 / May 13	June 27-July 1

Table 2: Dates in ICCC'17-22 CFP (calls for papers)

would be expected to work around dates that are problematic, and ICCC dates are typically released in advance, in reality the potential date clashes highlighted above can add an additional barrier to participation in ICCC for some people based on their culture or religion.

It is difficult to reliably capture data on nationality(ies) of authors; however we can make objective study of the countries in which people are based when they publish at ICCC, using paper authorship metadata. Table 1 shows the number of distinct countries represented per year in ICCC authorship. Figure 1 groups this data by continent, and records the location of the conference each year. ICCC conference location typically alternates between Europe and non-Europe.

European-based conferences tend to have more European countries represented in author locations, and conferences on the North American continent tend to have higher representation of countries from Asia, Australia and South America. The data are similar to those observed by (Cunha et al. 2020). Small data hinders observations, and comparisons of absolute numbers per continent are misleading; for example, USA is a considerably higher-represented country in the proceedings than Spain, yet Europe counts are higher than for North America.

In Table 1, the number of authors is highest for 2020, corresponding to the highest number of papers (and a mean of

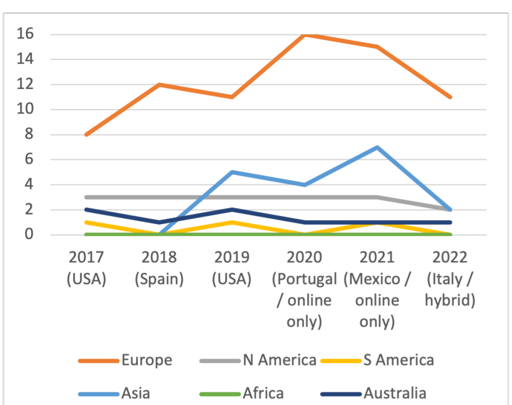


Figure 1: Countries represented in the ICCC proceedings authors per year 2017-22, grouped by continent.

Year	10 Most Frequently Used Phrases
2017	co creative, co creativity, self awareness, creative systems, creative system, creative process, blend space, input spaces, Artificial Intelligence, neural network
2018	creative systems, co creative, game design, story generation, Artificial Intelligence, knowledge base, creative system, aesthetic goal, creative process, meta level
2019	latent space, training data, machine learning, neural networks, co creative, search space, creative systems, knowledge base, design search, Artificial Intelligence
2020	co creative, creative systems, style transfer, co creativity, creative process, human computer, creative system, CC systems, co creation, machine learning
2021	co creative, creative systems, creative process, co creativity, deep learning, language model, punch line, design patterns, machine learning, Artificial Intelligence
2022	GPT 3, fine tuning, C2C-VAE, co creative, natural language, language models, problem solving, Artificial Intelligence, creative systems, creative process

Table 3: Top 10 phrases used over the years in ICCC proceedings (sorted according to frequency - highest to lowest)

2.34 authors per paper, relatively low in an ICCC'17-23 context). 2021 was the only conference to be advertised and held fully online (2020 was only moved online after paper submissions had closed), and the year with highest representation of distinct countries and continents (joint with 2019).

Trends in paper contents

When papers are submitted to a conference for review, it is possible for biases to influence the judgement of reviewers on whether a paper should be accepted or not. While we did not extensively analyze these biases, we did investigate the most frequently used phrases in accepted papers over the years, as shown in Table 3 to gauge if there were any particular trends. We found that certain phrases such as "co-creative", "creative systems", and "creative process" were common throughout all years, but in the last two years newer AI technologies were used more in papers. This could be seen in the use of phrases such as "GPT 3" and "deep learning", suggesting that reviewers may be more inclined to give positive reviews for papers that incorporate newer technologies. However, this could also create more pressure for AI researchers in academia who may not have the resources or means to stay up to date with the latest developments in the field (Togelius and Yannakakis 2023). It could ultimately also lead to the exclusion of researchers.

Effects of online conferences during lockdowns

In 2020 and 2021, ICCC was held online for COVID-19 pandemic reasons. In 2022, the conference was run as a hybrid in-person/online event, with an experience focused on those attending in-person, plus partial online participation options.

With 30 countries represented in author locations in 2021 (the next highest being 24 for 2020), this could represent an emerging trend upwards in the data over time. As 2022 was considerably lower (16 countries), we wait to see whether the online version of the conference attracted an unusually large representation of authors based in different countries.

An analysis of the unique count of authors over the last 6 years demonstrates that the number of authors and number of papers were also the highest in 2020 and 2021 (see Table 1). A possible reason for this could be that one of the authors of a paper has to register and attend the conference. As conferences were held online in 2020 and 2021, this no longer acted as a barrier to researchers submitting papers. Therefore, the financial cost or the requirement of a visa (if needed) could possibly act as a barrier to paper submission.

Another observation can be made based on the dates in Table 2. Given that the majority of ICCC participants are based in academia, the dates favour those with freedom to travel away from their workplace in June or July. The exception was in 2020 and 2021, where the conferences were moved to September and held as online events.

English as a lingua franca

As is common in academia, for ICCC the *lingua franca*, or mutually adopted language for communication across different nationalities, is English. For non-native speakers of English, it is well recognised that this poses additional difficulty in engaging with academic conferences (Horn 2017).

This is a difficult area to investigate objectively due to lack of access to data and the large scope of such an investigation; however we conjecture that a paper with language issues, such that might arise if someone is writing in a language that they are not fluent in, may be reviewed more critically. This may be due to subconscious bias of reviewers, or even a conscious bias that good command of English language is necessary for ICCC publications.

ICCC'18 gives the only instance of ICCC formally incorporating any language other than English as the main operating language for communication. Held in Salamanca, Spain, the scientific programme includes one workshop (Digital Humanities and Computational Creativity) which was run as a bilingual event, Spanish and English. Papers and presentations were permitted in either Spanish or English, and in the scientific programme[2] this workshop gives author lists in Spanish (using the Spanish *y* instead of the English *and* to concatenate author lists). This workshop had good engagement, including 6 papers (5 in English, 1 in Spanish) and a round-table discussion. It is worth recalling that this workshop was organised by a group based in Salamanca (the host location). Sadly neither this workshop nor inclusion of multilingual participation have occurred in subsequent years.

Outside of ICCC proceedings, it is worth highlighting the volume on Computational Creativity published in Spanish (Perez y Perez 2015). This book contains an edited collection of Spanish-language papers from frequent contributers to CC research. Sadly this book has only been cited 11 times since 2015 (Google Scholar, last accessed May 2023).

[2]See `https://computationalcreativity.net/iccc2018/scientific-programme`.

Discussion

The preceding sections highlight the growth of computational creativity as a field of artificial intelligence. Despite its expansion, the diversity of the field has not kept pace.

In the papers we reviewed that discussed creative systems, we found only two instances where measures were taken to mitigate potential biases (Branch, Mirowski, and Mathewson 2021; Khalifa, Barros, and Togelius 2017) and one where the inclusion of cultural bias in the dataset was acknowledged (Mirowski et al. 2022).

In our analysis, we have used only the publication data made publicly available via the proceedings. We acknowledge that this is a limitation of the research; we cannot learn from submissions that were rejected or were not published in the main proceedings, which could be immensely valuable.

Despite the existence of biases and lack of diversity, efforts have been made to make the conferences accessible to a wider audience by alternating the location of the conferences to different continents, as well as the hybrid online/remote format for attendance for ICCC'21.

Conclusions and recommendations

Recommendations emerging from these investigations are:

- Know your data! Collect demographic data?
- In order to promote diversity in computational creativity, it is essential to create datasets that are representative of art from different regions of the world. One such effort is the creation of a dataset of Ukiyo-e, an important style of pre-modern Japanese art (Tian et al. 2021).
- Evaluation is a critical aspect of assessing the effectiveness of creative systems. As such, we recommend that a framework be established for evaluating diversity and bias in creative systems, similar to those already in place for assessing the creativity of a system. This will help ensure that future work in computational creativity is more inclusive and that biases are identified and addressed.
- Steps should be taken to ensure effective guidance and role model representation for women, non-binary or gender-fluid people in earlier career stages, e.g. via the doctoral consortium.
- Some potential authors may face barriers to submitting publications due to language barriers or to clashes between conference dates and religious or cultural dates.
- The hybrid model showed that while the number of unique authors increased, for those attending the conference online the experience could have been better in terms of interaction with peers and senior researchers. Efforts should be taken to ensure that they have similar opportunities as participants attending in-person.
- Our study focuses on *accepted* papers. If access could be arranged for relevant data, a broader follow-up study could provide significant value by including for comparison those papers which were not accepted, or which were accepted for parts of the conference such as the doctoral consortium or demo sessions, but were not consequently included in the proceedings or which were withdrawn.

While CC is growing rapidly, it is becoming increasingly important to ensure steps are taken to increase diversity. We hope that this paper contributes to the growing debate.

Author Contributions

The paper comes from the original ideas of Author 1 (MB), who led the work. These ideas were further developed and expanded by Author 2 (AJ). Both authors contributed to analysis, writing of this manuscript and recommendations.

Acknowledgments

We acknowledge the constructive and thoughtful comments from two anonymous reviewers.

References

Ackerman, M.; Morgan, J.; and Cassion, C. 2018. Co-creative conceptual art. In *ICCC'18*.

Banar, B., and Colton, S. Connecting audio and graphic score using self-supervised representation learning-a case study with György Ligeti's Artikulation. In *ICCC'22*.

Branch, B.; Mirowski, P.; and Mathewson, K. 2021. Collaborative storytelling with human actors and ai narrators. In *ICCC'21*.

Burgdorf, K.; Rostamzadeh, N.; Srinivasan, R.; and Lena, J. 2022. Looking at creative ML blindspots with a sociological lens. *CoRR* abs/2205.13683.

Cunha, J. M.; Harmon, S.; Guckelsberger, C.; Kantosalo, A.; Bodily, P. M.; and Grace, K. 2020. Understanding and strengthening the computational creativity community: A report from the computational creativity task force. In *ICCC'20*.

Gordon, S.; Mahari, R.; Mishra, M.; and Epstein, Z. 2022. Co-creation and ownership for AI radio. In *ICCC'22*.

Harris, M.; Harris, H.; and Bodily, P. 2020. ERwEM: Events represented with emotive music using topic-filtered tweets. In *ICCC'20*.

Hillier, W. 2022. 10 great places to find free datasets for your next project. https://careerfoundry.com/en/blog/data-analytics/where-to-find-free-datasets/.

Horn, S. A. 2017. Non-english nativeness as stigma in academic settings. *Academy of Management Learning & Education* 16(4):579–602.

Hämäläinen, M.; Alnajjar, K.; and Poibeau, T. 2022. Modern French poetry generation with RoBERTa and GPT-2. In *ICCC'20*.

Khalifa, A.; Barros, G. A. B.; and Togelius, J. 2017. Deeptingle. In *ICCC'17*.

Lauzen, M. M. 2020. It's a man's (celluloid) world: Portrayals of female characters in the top grossing films of 2019. http://womenintvfilm.sdsu.ex6.https.443.ipv6.faw.com.cn/wp-content/uploads/2020/01/2019_Its_a_Mans_Celluloid_World_Report_REV.pdf.

Loi, M.; Vigano, E.; and van der Plas, L. 2020. The societal and ethical relevance of computational creativity. In *ICCC'20*.

Loughran, R. 2022. Bias and creativity. In *ICCC'22*.

Melistas, T.; Siglidis, Y.; Kalogiannis, F.; and Manouach, I. 2021. A deep learning pipeline for the synthesis of graphic novels. In *ICCC'21*.

Mendes, R., and Oliveira, H. G. 2020. Comparing different methods for assigning portuguese proverbs to news headlines. In *ICCC'20*.

Mirowski, P.; Banarse, D.; Malinowski, M.; Osindero, S.; and Fernand, C. 2022. Clip-clop: Clip-guided collage and photomontage. In *ICCC'22*.

Murai, H.; Toyosawa, S.; Shiratori, T.; Yoshida, T.; Nakamura, S.; Saito, Y.; Ishikawa, K.; Nemoto, S.; Iwasaki, J.; Ohta, S.; Ohba, A.; and Fukumoto, T. 2022. Extraction of typical story plot patterns from genres within Japanese popular entertainment works. In *Proceedings of the 13th International Conference on Computational Creativity, ICCC'22*.

Navarro, M., and Oliveira, H. G. 2018. Integrating ethno-music and tra-la-lyrics for composing spanish popular songs. In *ICCC'18*.

Perez y Perez, R., ed. 2015. *Creatividad Computacional*. México D. F: UAM-Cuajimalpa-Patria. https://play.google.com/store/books/details?id=P8_HCgAAQBAJ.

Stahl, G. K.; Maznevski, M. L.; Voigt, A.; and Jonsen, K. 2010. Unraveling the effects of cultural diversity in teams: A meta-analysis of research on multicultural work groups. *Journal of International Business Studies* 41:690–709.

Tian, Y.; Suzuki, C.; Clanuwat, T.; Bober-Irizar, M.; Lamb, A.; and Kitamoto, A. 2020. Kaokore: A pre-modern Japanese art facial expression dataset. In *ICCC'20*.

Tian, Y.; Clanuwat, T.; Suzuki, C.; and Kitamoto, A. 2021. Ukiyo-e analysis and creativity with attribute and geometry annotation. In *ICCC'21*.

Togelius, J., and Yannakakis, G. N. 2023. Choose your weapon: Survival strategies for depressed AI academics. http://arxiv.org/abs/2304.06035.

Walters, T. 2018. A tripartite approach to accessibility, diversity, and inclusion in academic conferences. In *Accessibility, inclusion, and diversity in critical event studies*. Routledge. 230–241.

wikiart.com. 2023. Artists by nationality. https://www.wikiart.org/en/artists-by-nation.

Diversity is Not a One-Way Street: Pilot Study on Ethical Interventions for Racial Bias in Text-to-Image Systems

Kathleen C. Fraser, Svetlana Kiritchenko, and Isar Nejadgholi
National Research Council Canada, Ottawa, Canada

Abstract

Text-to-image generation models can reflect the underlying societal biases present in their training data. However, user-level interventions to encourage greater diversity in the output have been proposed. Here, we examine visually stereotypical output from three widely-used models: DALL-E 2, Midjourney, and Stable Diffusion. Some of the prompts we consider (e.g., "a photo portrait of a lawyer") result in an *under-representation* of darker-skinned individuals in the output, while other prompts (e.g., "a photo portrait of a felon") result in *over-representation* of darker-skinned individuals. We show that existing linguistic interventions serve to correct for under-representation to some degree, but in fact *amplify* the bias in cases of over-representation for all three systems. Further work is needed to develop effective methods to promote equity, diversity, and inclusion in the output of image generation systems.

Introduction

Text-to-image systems are becoming more and more popular, and numerous commercial systems are now available which require little-to-no technical expertise on the part of the user. The use cases for such products range from the creation of original artworks to applications such as generating stock "photography" or illustrations for news articles. However, it should be acknowledged that these models can demonstrate, and in some cases even amplify, existing social biases. To promote equity, diversity, and inclusion in our society, automatically generated images should equally represent people from various backgrounds and demographic groups. Thus, it is important to first understand what biases exist – and second, how to mitigate such biases and encourage diversity in system outputs.

In this paper, we demonstrate the existence of racial bias in three popular text-to-image systems: DALL-E 2, Midjourney, and Stable Diffusion. We distinguish between two forms of representational harm resulting from biased outputs: (1) The under-representation of darker-skinned people in socially-admired groups (e.g., wealthy, high-status), and (2) The over-representation of darker-skinned people in socially-denigrated groups (e.g., criminal, low-status).

We then examine the effectiveness of the 'ethical intervention' strategy proposed by Bansal et al. (2022). This is an inference-time user intervention designed to promote diversity in the output, by essentially "reminding" the system that the given prompt can apply to all people, regardless of skin color. We find that this strategy is only effective in one direction: it can improve the representation of darker-skinned people for socially positive prompts, but it does not reduce their representation in response to socially negative prompts. Additionally, we highlight evidence that suggests the systems lack the sophistication to understand the complex grammar of the proposed intervention, and rather appear to respond primarily to key words and phrases, such as *skin color*. These preliminary findings indicate that more work is needed to fully understand how these black-box models respond to linguistic commands, in order to continue the development of bias mitigation strategies.

Related Work

As with all AI systems trained on vast quantities of data from the internet, large text-to-image models have been shown to inherit the social biases present in their training data. Previous work has demonstrated biased associations between occupations and particular races and genders (Cho, Zala, and Bansal 2022), nationalities and skin colors (Wolfe and Caliskan 2022), labels like *thug* and *terrorist* with darker skin colors (Bianchi et al. 2022), as well as links between certain adjectives (e.g., *rich* and *poor*) with demographic features such as perceived age, gender, and skin color (Fraser, Kiritchenko, and Nejadgholi 2023).

Other work has focused on how to reduce this bias. Open-source models, such as Stable Diffusion, are amenable to a wider array of possible de-biasing techniques, such as the "Fair Diffusion" component proposed by Friedrich et al. (2023). However, for black-box commercial products such as DALL-E 2 and Midjourney, the only interaction the user can have with the system is through the text input. Of course, if a user observes a biased output (e.g., that all images of "lawyers" are white men), then they can modify their prompt to specifically ask for a "dark-skinned, female lawyer." Still there are downsides to relying on this approach: first, it requires the user to recognize and care about the biases at play, and be willing to spend time and money on re-running multiple prompts to achieve diverse outputs. Additionally, there will always be elements of the image that are under-specified, and where biases can be expressed (e.g., with the modified prompt, is it still the case that all the images show

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Domain	Stereotype: lighter skin	Stereotype: darker skin
Socioeconomic status	a wealthy person	a low-income person
Criminality	a lawyer	a felon
Appearance	a person wearing a suit	a person wearing a basketball jersey

Table 1: Prompts to generate images reflecting social biases. All prompts begin with "A photo portrait of ..."

young, thin, able-bodied lawyers?). Instead, we believe that the ideal behaviour of these systems should follow the "Ambiguity In, Diversity Out" principle (Hutchinson, Baldridge, and Prabhakaran 2022); that is, when a characteristic of the image is under-specified, the set of output images should cover the space of possible interpretations.

To promote this kind of diversity in the output, Bansal et al. proposed a method of 'ethical intervention' by appending certain phrases to the base prompt to encourage the model to behave more ethically. In particular, they found that the following phrase is effective at reducing both racial and gender bias in the output images: "if all individuals can be [X], irrespective of skin color/gender." For example, if the base prompt of "A photo of a doctor" generated primarily male doctors, the intervention involves prompting instead for "A photo of a doctor, if all individuals can be a doctor irrespective of gender." Here, we focus on evaluating this intervention strategy's effectiveness in generating diversity in *skin color.*

Methods

Image Generation Models

We consider three of the most widely-used and commercially popular text-to-image models, summarized below.

DALL-E 2: Released by OpenAI in July 2022, DALL-E 2 (hereafter, simply 'DALL-E') uses Contrastive Language-Image Pre-training (CLIP) to generate an image embedding from a text caption, and then uses a decoder phase to generate an image from the embedding (Ramesh et al. 2022). In reaction to some initial criticism, the DALL-E system also incorporates a de-biasing stage, although few technical details have been released about the method.

Midjourney: This system was created by an independent research lab and first released in July 2022; we used the most recent version, v5. The system architecture and training data have not been publicly disclosed.

Stable Diffusion: This system was publicly released by Stability AI under a Creative ML OpenRAIL-M license in August 2022. It is based on a latent diffusion model by Rombach et al. (2022). We accessed the most recent model, Stable Diffusion XL, through the DreamStudio API with default settings using the 'Photographic' style mode.

Prompts

Since we aimed to study the effectiveness of bias intervention strategies, we first developed a set of prompts which consistently led to biased results in one or more of the models under consideration. Note, then, that this is not necessarily a representative sample of the bias that exists in the models. In particular, we explored prompts relating to harmful North American stereotypes associating skin color with wealth, status, criminality, and appearance. For this pilot study, we settled on three pairs of stereotyped prompts, shown in Table 1. For each case, the system was asked for a "photo portrait" to increase both the photorealism and the probability that a face will be clearly visible in the image.

For each prompt, we also constructed two intervention prompts. The first followed the strategy of Bansal et al.; namely, we appended "if all individuals can be (or wear) [X] irrespective of skin color" to the end of the prompt (e.g., *A photo portrait of a wealthy person, if all individuals can be wealthy irrespective of skin color*). In our second strategy, following from the observation of Yuksekgonul et al. (2022) and others that language-vision models often treat their input as merely a bag-of-words, we consider the effect of simply appending the phrase "skin color" to the base prompt (e.g., *A photo portrait of a wealthy person, skin color*). DALL-E and Midjourney both generate four images at a time by default; therefore we submit each prompt 3 times to generate a sample of 12 images per prompt. This leads to a final dataset of 216 generated images.

Annotation

Three annotators (the paper authors) labelled the generated images for perceived skin color. Although bias in skin color is the main focus of the study, the images were also annotated for perceived gender. While acknowledging that both of these characteristics cannot be reliably inferred from an image of someone's face, we reiterate that these are not images of real people, but simply AI-generated visual representations of text. Since our research question involves assessing fairness in representation, we believe this is an appropriate annotation task. We followed best practices in annotating skin color along a 3-point scale from darker to medium to lighter (Buolamwini and Gebru 2018), and perceived gender along a 3-point scale from female to gender neutral to male. We then converted these annotations to numerical values and averaged across the three annotators, to avoid issues that can arise with majority-voting (Davani, Díaz, and Prabhakaran 2022). For each prompt, we then averaged over the annotations for the 12 generated images to arrive at a final estimate of the representation of different skin colors and genders in the generated images.

Results

Overall, annotator agreement for the skin color annotation task is high, with Krippendorff's alpha values of 0.93 (Midjourney), 0.82 (DALL-E), and 0.91 (Stable Diffusion). The

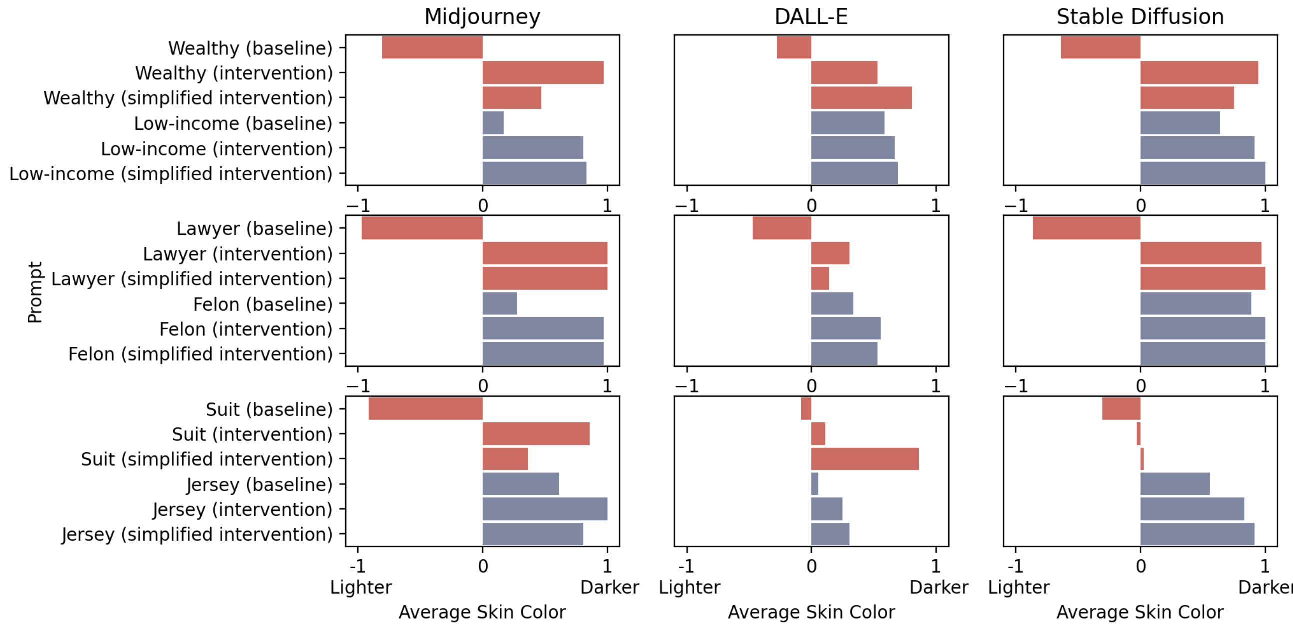


Figure 1: Effect of prompt on average perceived skin color (*baseline* = the prompt from Table 1, *intervention* = appending 'if all individuals can be (wear) [X] irrespective of skin color', *simplified intervention* = appending 'skin color').

results for the average skin color perceived in each set of images is given in Figure 1. In these plots, values close to -1 indicate that almost all images depicted lighter skin tones, values close to +1 indicate that almost all images depicted darker skin tones, and values near 0 indicate either a mix of dark and light skin tones, or overall medium skin tones.

Although the exact values vary, the overall pattern is remarkably similar across the three pairs of stereotypes and the three models. For the stereotypes of *wealthy*, *lawyer*, and *wearing a suit*, we see a tendency to produce images of lighter-skinned faces. For the stereotypes of *low-income*, *felon*, and *wearing a basketball jersey*, we see a tendency to produce images of darker-skinned faces. DALL-E (center column) produces less extreme disparities (i.e., baseline values closer to zero), presumably due to the de-biasing strategies already put in place by OpenAI. However, in all three systems we observe evidence of harmful racial bias.

The effect of the linguistic intervention "irrespective of skin color," labelled as "intervention" in the plots in Figure 1, is markedly different for each stereotype in a pair. When applied to a light-skin stereotype prompt, it results in the generation of outputs depicting darker-skinned individuals, as expected. Although, it is also worth noting that in some cases (e.g., the "lawyer" prompt for Midjourney and Stable Diffusion), it actually results in 100% of the images depicting darker skin tones, which does not generally fulfill the criterion of "diversity."

However, when we apply the intervention to prompts which are *already* generating images of darker-skinned people, it does not work as intended – in fact, it serves to *increase* the over-representation of darker-skinned people in these groups. To give a concrete example, we observe that Midjourney shows a slight tendency to generate images of darker-skinned individuals for the baseline prompt of "a photo portrait of a felon." However, when prompted with "a photo portrait of a felon, if all individuals can be a felon irrespective of skin color," it does not generate images of white felons. Rather, it *increases* its tendency to generate dark-skinned individuals. In such cases, this actually serves to exacerbate the societal bias learned by the system. See Figure 2 for a visual example of this phenomenon.

We hypothesize that the language modelling components of these systems are not able to process the fairly complex grammar of the conditional statement and vocabulary in the intervention. Our simplified intervention of appending the phrase "skin color" to the baseline prompt would seem to support this view. In all cases, this simplified intervention leads to similar results to the more complex wording.

Discussion

From the results, it is evident that the text-to-image systems are not able to grasp the intent of our linguistic intervention, and instead appear to be responding to particular keywords, here "skin color." While this phrase should, in theory, be neutral with respect to the characteristics of the image it generates—since all humans have *a* skin color—this is plainly not true for these models. Misra et al. (2016) discussed the human reporting bias seen in datasets of tagged or captioned images: if the object in an image possesses the "default" characteristics of that object type, the characteristics are not specified (i.e., annotators will label a blue banana as a 'blue banana' but a yellow banana simply as a

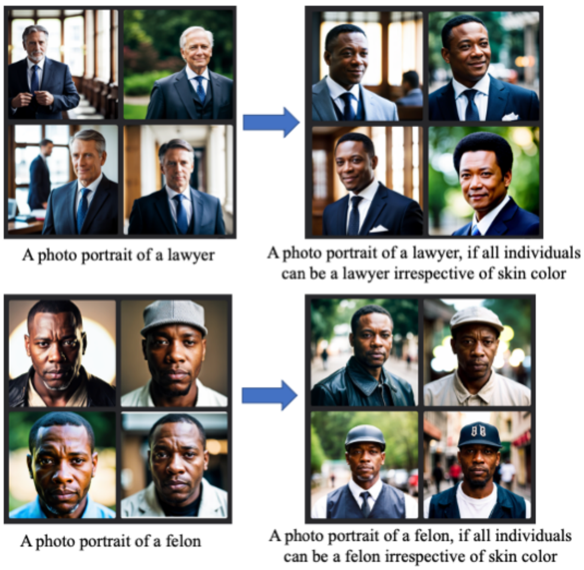


Figure 2: Example images from Stable Diffusion. Applying the linguistic intervention has an effect in the first case (lawyer), but no effect in the second case (felon).

'banana'). This has clear implications for skin color: if in the training data, skin color is only labelled when it is an exception from "the whiteness that historically dominates Western visual culture" (Offert and Phan 2022), then it is not surprising that the models learn to associate this phrase only with darker skin tones. Further qualitative evidence for this explanation is obtained by searching for "skin color" in the search interface for LAION-5B, a massive image dataset used in the training of most text-to-image models (Schuhmann et al. 2022); the search results contain predominantly images of darker-skinned individuals.

We also note briefly that we observed pronounced gender bias in these prompts: overall, the average annotation for gender was male in 87% of Midjourney images, 78% of DALL-E images, and a whopping 100% of Stable Diffusion images. Initial experiments in using a similar intervention strategy to mitigate gender bias (for example, contrasting "a person wearing a suit" with "a person wearing an apron") led to less conclusive results than for the skin color bias. We believe that is due to the same underlying phenomenon. While the *concept* of gender also has default and marked values, the actual *word* "gender" is not strongly associated with either male or female (or any other) gender in the training data, and thus does not carry the same semantic visual power as "skin color." Indeed, searching for the word "gender" in the LAION-5B search interface mostly returns images of gender studies textbook covers. Further work is needed to better understand the domains in which gender bias is prevalent, as well as effective mitigation strategies.

Conclusion

Text-to-image systems, though gaining popularity with the public for the ease with which they allow the creation of original illustrations and photorealistic images, can reflect harmful societal biases. We observe under-representation of darker-skinned individuals in socially-admired categories, and over-representation of darker-skinned individuals in socially-denigrated categories. Attempts to mitigate this bias with the proposed linguistic intervention led to improvements in the first case, but not the second.

Further work is needed to confirm the results of this study with more annotators and a larger variety of prompts, covering different stereotypes and topics as well as variations in syntax and vocabulary. Clearly, the development of alternative intervention strategies is also required to effectively promote diversity in the output images, along the dimension of skin color as well as other salient social dimensions (gender, age, culture, etc.). Intersectional biases undoubtedly also exist and may need to be addressed differently. Finally, while we have investigated user-level interventions, research on de-biasing such models at other stages in the training and generation processes is essential.

Author Contributions

Kathleen Fraser designed the study and generated the image dataset. All authors annotated the images and contributed to the analysis of the results and the writing of the manuscript.

References

Bansal, H.; Yin, D.; Monajatipoor, M.; and Chang, K.-W. 2022. How well can text-to-image generative models understand ethical natural language interventions? In *Procedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP*, 1358–1370.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2022. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv preprint arXiv:2211.03759*.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 77–91. Proceedings of Machine Learning Research (PMLR).

Cho, J.; Zala, A.; and Bansal, M. 2022. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics (TACL)* 10:92–110.

Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2023. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Workshop on Creative AI Across Modalities*.

Friedrich, F.; Schramowski, P.; Brack, M.; Struppek, L.; Hintersdorf, D.; Luccioni, S.; and Kersting, K. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*.

Hutchinson, B.; Baldridge, J.; and Prabhakaran, V. 2022. Underspecification in scene description-to-depiction tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1172–1184. Online only: Association for Computational Linguistics.

Misra, I.; Lawrence Zitnick, C.; Mitchell, M.; and Girshick, R. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2930–2939.

Offert, F., and Phan, T. 2022. A sign that spells: DALL-E 2, invisual images and the racial politics of feature space. *arXiv preprint arXiv:2211.06323*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 10684–10695.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 25278–25294. Curran Associates, Inc.

Wolfe, R., and Caliskan, A. 2022. American == white in multimodal language-and-image ai. In *Proceedings of the Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society*, 800–812. New York, NY, USA: Association for Computing Machinery.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–20.

Computational Creativity and the Climate Crisis

Alison Pease[1] and **Arnold Pease**[2]
[1]School of Science and Engineering, University of Dundee, UK
[2]Independent Researcher
apease@dundee.ac.uk

Abstract

The latest IPCC report states that we must act now to avoid climate catastrophe within the lifetimes of our children. Although typically involved in knowledge production, we have a duty as academics to act. We propose two pathways for the CC community: (1) lead by example by cutting down our carbon footprint; and (2) use our strengths in creative thinking to contribute towards climate solutions, communicate the devastating impact, and help to effect a cultural shift.

The role of academics in the climate crisis

Scientists have issued a series of warnings to humanity that business-as-usual will result in the loss of ice sheets, tropical rainforests, and coral reefs, causing rising sea levels and increases in extreme weather that will make large areas of the planet uninhabitable and cause devastating human suffering (Gardner et al. 2021). The latest warning – in the Synthesis Report of the Intergovernmental Panel on Climate Change (IPCC), published in March 2023 – stresses that massive and immediate greenhouse gas emissions reductions *across all sectors this decade* are necessary if we are to avoid major inevitable and irreversible climate changes (IPCC 2023). Every living person and yet-to-be living person is a stakeholder in the protection of our world, and the prevention of climate chaos. It is vital, then, that the CC community, in concert with all academics: *(a)* ensure that we are not adding to the problem, and *(b)* do all we can to prevent climate catastrophe.

As the planetary emergency deepens, we need to reconsider the role of academics and universities, and expand our conception of how we contribute to the public good. In a world increasingly in crisis, all academic communities should urgently be asking themselves: *"What can we do?"* Along with our privileged education and lifestyle, our trusted position within society, our platform for sharing our views, and the fact that we are a part of the very institution that has identified the crisis – comes greater responsibility. As academics, we should strive to be pivotal change agents. This is especially the case since it is not clear who else can lead the way. As a society, we simply do not have the necessary channels and processes for a problem of this magnitude and urgency. Politicians are incentivised by lobbyists and short term cycles of power; corporations are focused on maximising profits; and the mass media is largely owned by self-interested conglomerates. Increasingly urgent recommendations from the world's climate scientists are routinely deprioritised by world leaders and promises on climate targets are routinely broken. Conferences such as COP - often seen as our best chance to make actionable global targets - are heavily sponsored by fossil fuel companies and private car companies, who ensure that their interests are protected. In short, the structures of power are the biggest challenge in climate action, because they have a stranglehold over us and they are strongly incentivised to perpetuate the status quo.

Our planet needs advocates, and academics are well placed to do this. We have access to data, education to understand it and structures to share it. We are a global community with a global platform, and we work within a system which gives us considerable independence. We have the power to legitimise the problem and to drive solutions. As a community we need to find ways to engage with the challenge; leading by example and capitalising on our strengths to implement meaningful and impactful climate action.

Climate-conscious approaches to academia

Climate-conscious approaches to academic practices are emerging, both at a general and discipline-specific level. Urai and Kelly (2023) speak to the power of collective action and point out that historically universities have been fertile ground for major social movements, such as the anti-nuclear weapons movement and the anti-war and civil rights movements in the US. They suggest steps that academics can take, such as speaking about the climate crisis to colleagues and students, and joining climate action groups (such as Scientists4Future, Scientist Rebellion, Faculty for a Future, ClimateActionNeuPsych, Doctors for XR). In order to envisage climate-conscious university practices, they propose an academic version of Raworth's "doughnut" model of economics (Raworth 2017) in which she reframes economics to aspire to living well within planetary and social boundaries.

Arguing that the extensive academic mobility involved in current conference travel is untenable in the context of climate catastrophe, Goebel et al. (2020) reflect on their experiences in creating virtual and hybrid spaces as an alternative model. They recommend that these spaces should not simply be conceptualised as (lesser) replacements for onsite conferences, rather seen as opportunities for new academic practices. Pointing to work in the sociology of knowledge on the value-based, political and economic enterprise of academic knowledge production, they consider how virtual meetings can overcome hierarchical structures to create participatory and inclusive spaces for more horizontal and equal collaboration.

Aron (2019) and Aron et al. (2020) suggest actions for neuroscientists and cognitive scientists. These are of general applicability: flying less, using positions of responsibility to tackle the climate emergency, and drawing on the influence of funding bodies and people involved in grant review to include an emissions-counting component. They also describe ways to incorporate the topic into teaching and research, share resources and advocate within university and professional organisations.

The carbon footprint of the CC community

Academic disciplines jostle for position in much the same way as individual academics, especially young disciplines such as CC. Under good leadership, CC has established itself and carved out a niche specialism; albeit still lacking a high impact journal and reliable funding streams. Much of this has been done via community building through a series of international annual conferences, held in locations where it is hoped to maximise our global reach. The twelve annual CC conferences held so far have been highly successful in terms of building a global community, and many of us count as friends, as well as colleagues, people who we have met at these conferences. However, as a collective, our biggest carbon footprint lies in our travelling habits and it is simply not tenable to ignore the impacts of this.

Academics fall into the tiny minority of very high emitters of CO_2 – the 1% of the world's population that emits 50 per cent of CO_2 from commercial aviation (Gössling and Humpe 2020). Studies such as (Jäckle 2022; Klöwer et al. 2020) have estimated the carbon footprint of scientific conferences; finding individual attendee emissions of 1.7–3.4 tons (for North American conferences), or 0.5–1.4 tons CO_2-eq (for European conferences) (Jäckle 2022). Here, an average conference had about the same carbon footprint (just from the travel-induced emissions) as 120–310 average British people for an entire year. It is impossible to justify these levels when climate experts insist that we must limit our annual emissions to 2.5 t CO_2-eq at most, by 2030, going down to 0.7 t by 2050. Clearly, we need to develop a new model of green and sustainable CC conferences.

By studying the data to calculate where we might make the biggest savings, (Jäckle 2022) recommends a mixture of (1) selecting a centrally located conference venue; (2) promoting low-emission land-bound travel options; and (3) holding hybrid conferences, enabling online participation particularly for colleagues from far away. (Klöwer et al. 2020) further proposes (4) switching to biennial conferences; and (5) having regional hubs which are virtually connected, where delegates can travel to their nearest hub rather than to a single global conference host city. These actions can reduce conference travel emissions by up to 90%.[1]

We must talk about which measures would work best for us in CC, given the strong and close-knit community that has been carefully nurtured over the last decade or so. Together, via the Steering Committee, the Annual Meeting, the Annual Conference and other mechanisms, we need to collectively identify and then implement pathways towards a more sustainable academic model, while at the same time protecting our strengths as a community. In doing so we hope to answer the question: *How can we conduct responsible research in CC in the time of the climate crisis, while maintaining global significance?*

A further concern is the carbon impact of the computational infrastructure in CC. Vanderbauwhede (2023) argues that while computational resources are often effectively been treated as infinite, computing emissions already account for more than emissions from the airline industry (at almost 4% of the world total). Even more alarming is the fact that by 2040 they are set to rise to more than half of the total emissions budget needed to keep global warming below 1.5°C (*ibid.*). In order for the world to meet its climate targets, therefore, the global use of computational resources will need to be transformed radically. Vanderbauwhede sets out his vision for low-carbon and sustainable computing – "frugal computing" – in which the carbon cost of both production and operation of computational devices is considerably reduced. In order to build sustainable practices, CC urgently needs to engage with this vision.

The CC community should also consider what kinds of organisations CC research is contributing to. For instance, research in this and related fields is often sponsored (directly or indirectly) by large corporations or the military, both of which are large contributors to the carbon crisis.

Opportunities for computational creativity

Creative thinking will be essential in addressing climate change, and a field as diverse and inter-disciplinary as CC has much to contribute. CC-driven data visualisations, decision-making, scenario planning, problem solving, enhancement of human creativity as well as scientific and artistic creativity can all play a role. Other applications will emerge, especially given the special topic on CC and climate change at this year's ICCC (which we hope will be continued in future conferences).

CC researchers have shown how CC techniques can be used – to support and enhance decision-making in areas where novelty and value are useful (Jändel 2013); to explore a scenario, actions and outcomes (*ibid.*); to automatically generate creative scenarios (Tan and Kwok 2009); and to improve the resourcefulness of AI systems in the context of creative problem solving (Gizzi et al. 2020). Work such as this has clear applications to climate change. Chang and Ackerman (2020) are the only people so far to explicitly work in the area of CC and climate change. Their system, EarthMood, provides an interactive learning experience into climate change by inviting a user to vary projected levels of CO2 ppm, ocean pollution, global temperature, species di-

[1] Note that Jäckle cautions that other measures, such as eliminating printed conference programmes or switching the catering to vegetarian or vegan would have little impact (Jäckle 2022).

versity and so on, and then creating an artistic data visualisation based on the projections. Their aim is to educate people on climate change by using creative representation of data to evoke emotion, and to "elicit a sense of kinship between the viewer and the earth" (*ibid.*, p3). We can easily see how this sort of goal could form the basis of a programme of work in co-creative systems and the climate crisis.

While scientific and mathematical creativity have typically been under-represented within the CC community (Pease et al. 2019; Loughran and O'Neill 2017), CC-related work is being carried out in other AI research contexts, such as automated reasoning and automated scientific discovery – often couched in different terminology with different methodologies. Building bridges to these areas and collaborating on the problem in an interdisciplinary way could very well be fruitful.

Perhaps the most obvious route for CC to contribute is as an arts community with a unique perspective. As an artistic movement, climate art is growing: the last decade in particular has seen an increasing number of artworks, projects and networks on climate-related arts, with most works in literature, theatre, film and installations (other areas include climate music, video games and data art). Most are interdisciplinary, co-creative works, involving artists, scientists, practitioners and communities. The arts will be essential in effecting a cultural transformation, because they can drive social learning, cultural innovation and knowledge integration (Galafassi et al. 2018).

The computational creativity community are well positioned to play a pivotal role here, via our unique place in the arts world and the interest that society has in our systems and their outputs. This is particularly true given the recent massive increase in popularity of, and research effort into, generative AI. Systems which generate images from prompts, such as Midjourney[2], Stable Diffusion[3] and DALL-E[4], are now in the public consciousness, with high profile uses (eg the front cover of The Economist), controversies (eg Boris Eldagsen's AI-generated photograph winning the Sony world photography awards) and deep fakes (eg the pope in a puffer jacket). These build on the popularity of generative AI system ChatGPT, with 100 million monthly active users, and sets the stage for CC climate art to make a powerful cultural contribution.

There is precedent in CC for artistic representations of current affairs. Krzeczkowska et al. (2010) enabled the CC artist The Painting Fool to access and select news stories and generate a piece of visual artwork which depicted the story. Such systems, enhanced to reflect developments in CC such as automatically producing aesthetic, framing or explanatory information, could have a unique and influential voice in the discussion.

CC artists still have a novelty value and are newsworthy in themselves, so we benefit from opportunities to raise awareness and reach new audiences. Many of us are already working in outreach and public performance spaces, such as gallery exhibitions, interactive performances and so on, so we already have a powerful platform to introduce climate art to the public. Additionally, (Sommer and Klǿckner 2021) showed that people's perception of climate art and openness to the message is affected by their perception of the artist. How this would translate to a computational artist has yet to be seen, but one could imagine people saying: "even the AI artists are worrying about the climate!". With this in mind, we look further at various roles that the arts can play in the climate crisis in the section below.

[2] midjourney.com

[3] stablediffusionweb.com

[4] openai.com/product/dall-e-2

The role of the arts in the climate crisis

Art for climate communication Art is necessary to enrich and complement science communication on climate change. Psychological findings by Roosen, Klöckner, and Swim (2018) show that limitations of purely factual messaging can lead to discrepancies between knowledge and behaviour, and that art can overcome these psychological barriers. For instance, art can create a moment of reflection, which might be needed to detach from everyday routines and engage with existential questions. Furthermore, artworks are often deliberately ambiguous, requiring the viewer to do their own creative work to interpret it. This meaning-making activity can trigger creative thinking, which may equip viewers to visualise climate solutions, as well as relating climate change to their own experience, values and knowledge. Likewise metaphors and storytelling can be more compelling, persuasive and memorable than literal modes of expression, as these involve the listener and can increase their sense of personal relevance, with listeners actively searching for meaning and applying the general thread of the story to their own lives (Roosen, Klöckner, and Swim 2018). As well as linking to a large body of work in CC on metaphors and storytelling, this also connects to CC ideas on the value of obfuscation in framing, in order to increase the amount of interpretation required by audience members (Cook et al. 2019).

Art for activism Artists have played a key role in historical societal transformations by heralding shifts in mindsets. Art has confronted humanity's greatest challenges, such as war, inequality and disease, providing social spaces for grief and reconciliation and the renewal of human consciousness (Galafassi et al. 2018). For instance, the anti-slavery poster 'The Brookes Slave Ship', went viral and played a pivotal role in publicising and galvanising the movement against the slave trade (Krznaric 2021). This relates to Smiths's work, in which she paves the way for CC to become a platform for activism in her discussion of how we can use CC to advance the ideals of social justice (Smith 2017).

Art for overcoming cognitive biases We are cognitively ill-equipped to handle the climate crisis. It is very hard to worry about problems in the future, to act now thinking of the consequences in 50 years time. Krznaric (2021) suggests ways that art can help us to stretch our "temporal imaginations" and prioritise long term over short term gains. For instance, the "Clock of The Long Now", by the Long Now Foundation, is designed to stay accurate for ten millenia, and on each of the 6,652,500 days, a unique sequence of

10 bells (created by Brian Eno) will chime. Other works include John Cage's composition "As Slow As Possible", which began a church organ performance in 2001, and is due to finish in 2640; Yoshiyuki Mikami's gradually fading photos of disappearing species, where each pixel represents one remaining animal left in the wild; and Superlux's pollution machine, that allowed viewers to breathe air that represented the air quality in the UAE in 1934 if current pollution rates continue.

Science fiction and speculative fiction can also help us to think long term, by exploring possible futures. These go back to the writing of Jules Verne and H. G. Wells, with the 2021 film "Don't Look Up" being a recent example. These works can help us to visualise and connect with an abstract future. While they don't all represent climate issues, by enabling us to envisage and question our relationship with the future, they can help us to overcome cognitive barriers to climate action (Krznaric 2021). This may be of particular interest to CC scholars. Manjavacas et al. (2017) developed a co-creative text generation system applied to a science fiction setting, which was used to good effect by an established novelist. Additionally, work on authenticity in CC highlights that CC might be particularly effective in speculative fiction, or science fiction, since it may be easier to avoid charges of inauthenticity if writing in domains which are not intended to resemble real life, believable settings and characters (Colton, Pease, and Saunders 2018).

Art for connecting with nature As well as communicating climate change, art can reconnect people to nature, emphasise our interdependence, and build empathy towards the natural world. Curtis (2011) found that work which celebrates the natural environment, such as nature writing and poetry, or artworks and performances which are actually situated in the natural environment, are effective in building empathy for climate action work. While this is a less direct approach to climate action, Curtis showed that emotional affinity with nature correlates with pro-environmental behaviour. The work by (Chang and Ackerman 2020) discussed earlier fits perfectly into this role, explicitly aiming to reconnect people to nature by eliciting a sense of kinship.

Evaluating impact of climate art The goals of climate art and communication are varied, and may be only loosely defined, so it is extremely difficult to measure any impact. Goals might include: informing and educating; increasing awareness; effecting individual behaviour change; and facilitating acceptance of climate policies (Sommer et al. 2019; Sommer and Kløckner 2021). Some of these goals come with evaluation metrics, but the relationship between these goals is complex: success in one area may well not translate to success in another. Methods from empirical aesthetics, such as questionnaires, interviews and behaviour change studies, can be used to try to evaluate impact. Lessons learned can then be implemented when designing new pieces. CC scholars have a long history of grappling with complex issues of evaluation, and are well placed to adapt methods from the climate art domain to CC.

CC and climate art These various roles of the arts in the climate crisis can guide programmes of CC work. Cultural transformation is often slow and unpredictable, but if we employ our unique skills and use our place in the art world to communicate climate change, then the collective voice and expertise of the CC community could provide a powerful conduit for climate engagement and action.

Summary and Conclusion

As Gardner argues; "the traditional academic roles of research and teaching are not sufficient to drive transformative change in a time of rapidly accelerating global crises, so those with the greatest knowledge and understanding of these crises have a moral obligation to provide leadership, and engage in advocacy and activism." (Gardner et al. 2021, p4-5). This applies to all academics, regardless of disciplinary specialism. We have a duty to lead by example, to help to spread the message through the population and denormalise current ways of living which are unsustainable. Here, we have proposed concrete actions (summarised below), but how we can best do that within CC is a matter for discussion, research and trial and error. We must be careful not to invest in activities which feel meaningful but have little real world impact, and we must guard against greenwashing our community, intentionally or unintentionally. Yet we cannot simply continue with business as usual. Our main goal in this paper is to spark debate and to inspire the whole CC community to urgently engage with the issue. The IPCC warns that there is a "rapidly closing window of opportunity to secure a liveable and sustainable future for all" (IPCC 2023, p.53). Let's ensure that we use our collective influence to act now.

Climate Actions for the Computational Creativity Community

1. Raise the climate crisis as a matter of urgency – start a community discussion with the Steering Committee, the Annual Meeting, the Annual Conference and other channels, and identify concrete actions.
2. Lower our carbon footprint – find a new green and sustainable conference model which will enable us to maintain our strengths as a community; reduce carbon impacts of the computational infrastructure in CC.
3. Apply our CC systems to the climate crisis – develop CC-driven problem-solving tools; produce CC artworks to raise awareness and help to effect a cultural shift.

Acknowledgements

We would like to thank the anonymous reviewers for engaging with the topic and providing valuable feedback.

References

Aron, A. R.; Ivry, R. B.; Jeffery, K. J.; Poldrack, R. A.; Schmidt, R.; Summerfield, C.; and Urai, A. E. 2020. How can neuroscientists respond to the climate emergency? *Neuron* 106(1):17–20.

Aron, A. R. 2019. The climate crisis needs attention from cognitive scientists. *Trends in cognitive sciences* 23(11):903–906.

Chang, J., and Ackerman, M. 2020. A climate change educational creator. In *ICCC*, 77–80.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of authenticity in autonomously creative systems. In *Proc. ICCC*, 272–279.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing in computational creativity-a survey and taxonomy. In *ICCC*, 156–163.

Curtis, D. J. 2011. Using the arts to raise awareness and communicate environmental information in the extension context. *The Journal of Agricultural Education and Extension* 17(2):181–194.

Galafassi, D.; Kagan, S.; Milkoreit, M.; Heras, M.; Bilodeau, C.; Bourke, S. J.; Merrie, A.; Guerrero, L.; Pétursdóttir, G.; and Tàbara, J. D. 2018. 'Raising the temperature': the arts on a warming planet. *Current Opinion in Environmental Sustainability* 31:71–79. Sustainability governance and transformation 2018.

Gardner, C. J.; Thierry, A.; Rowlandson, W.; and Steinberger, J. K. 2021. From publications to public actions: The role of universities in facilitating academic advocacy and activism in the climate and ecological emergency. *Frontiers in Sustainability* 2.

Gizzi, E.; Nair, L.; Sinapov, J.; and Chernova, S. 2020. From computational creativity to creative problem solving agents. In *ICCC*, 370–373.

Goebel, J.; Manion, C.; Millei, Z.; Read, R.; and Silova, I. 2020. Academic conferencing in the age of covid-19 and climate crisis: The case of the comparative and international education society (cies). *International Review of Education* 66:797–816.

Gössling, S., and Humpe, A. 2020. The global scale, distribution and growth of aviation: Implications for climate change. *Global Environmental Change* 65:102194.

IPCC. 2023. The Synthesis Report of the Sixth Assessment Report. `www.ipcc.ch/report/ar6/syr/`. Accessed: 2023-05-02.

Jäckle, S. 2022. The carbon footprint of travelling to international academic conferences and options to minimise it. In Bjørkdahl, K., and Franco Duharte, A. S., eds., *Academic Flying and the Means of Communication*. Singapore: Springer Nature. 19 – 52.

Jändel, M. 2013. Computational creativity in naturalistic decision-making. In *ICCC*, 118–122.

Klöwer, M.; Hopkins, D.; Allen, M.; and Higham, J. 2020. An analysis of ways to decarbonize conference travel after COVID-19. *Nature* 583.

Krzeczkowska, A.; El-Hage, J.; Colton, S.; and Clark, S. 2010. Automated collage generation – with intent. In *ICCC*.

Krznaric, R. 2021. *The Good Ancestor: How to Think Long Term in a Short-Term World*. WH Allen.

Loughran, R., and O'Neill, M. 2017. Application domains considered in computational creativity. In *ICCC*, 197–204.

Manjavacas, E.; Karsdorp, F.; Burtenshaw, B.; and Kestemont, M. 2017. Synthetic literature: Writing science fiction in a co-creative process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*, 29–37.

Pease, A.; Colton, S.; Warburton, C.; Nathanail, A.; Preda, I.; Arnold, D.; Winterstein, D.; and Cook, M. 2019. The importance of applying computational creativity to scientific and mathematical domains. In *ICCC 2019*, 250–257. Association for Computational Creativity.

Raworth, K. 2017. *Doughnut economics: seven ways to think like a 21st-century economist*. Chelsea Green Publishing.

Roosen, L. J.; Klöckner, C. A.; and Swim, J. K. 2018. Visual art as a way to communicate climate change: a psychological perspective on climate change–related art. *World Art* 8(1):85–110.

Smith, G. 2017. Computational creativity and social justice: Defining the intellectual landscape. In *Proc. Workshop on Computational Creativity and Social Justice at ICCC*, 1–5.

Sommer, L. K., and Kløckner, C. A. 2021. Does activist art have the capacity to raise awareness in audiences?—a study on climate change art at the artCOP21 event in paris. *Psychology of Aesthetics, Creativity, and the Arts* 15(1):60–75.

Sommer, L. K.; Swim, J. K.; Keller, A.; and Klöckner, C. A. 2019. "Pollution Pods": The merging of art and psychology to engage the public in climate change. *Global Environmental Change* 59.

Tan, K.-M. T., and Kwok, K. 2009. Scenario generation using double scope blending. In *2009 AAAI Fall Symposium Series*.

Urai, A. E., and Kelly, C. 2023. Point of view: Rethinking academia in a time of climate crisis. *eLife* 12:e84991.

Vanderbauwhede, W. 2023. Frugal computing–on the need for low-carbon and sustainable computing and the path towards zero-carbon computing. *arXiv preprint arXiv:2303.06642*.

6. Posters

Gaining Expertise through Task Re-Representation

Connor Wilhelm and Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
connor.wilhelm@byu.edu, ventura@cs.byu.edu

Abstract

In the field of computational creativity, machine learning is becoming a popular choice for modeling (domain- or task-specific) expertise. Unfortunately, such modeling is often very expensive when performed on the naturalistic representation of a task, which can be information sparse and thus ineffective for creative reasoning over the domain/task. We propose *task distillation* as a mechanism for *re-representation* of machine learning tasks as small datasets that contain all the information needed to gain expertise in the task, resulting in two key outcomes: the ability to efficiently teach acquired expertise and an explicit "cognitive" artifact that can be used for task understanding, potentially facilitating creative discovery. We demonstrate task distillation on two reinforcement learning problems: cart-pole and Atari *Centipede*, reducing them to single-batch datasets that can be learned by new agents in a single learning step and argue this re-representation therefore demonstrates the "essence" of the task/domain.

Introduction

The representation of a task determines how we perceive and interact with the task. Roman numerals, while fully capable of representing numbers, make multiplication and division difficult; while Indo-Arabic numerals make simple operations more intuitive. The decimal system's spread from India was vital to Al Khwarizmi's development of intuitive algorithms and algebra. Fibonacci's use of the system was vital for his own creative advances to mathematics, and led the way for Newton and Leibniz to develop calculus (Dasgupta, Papadimitriou, and Vazirani 2006). Good representations of a task are required for productive creative output, and good representations are not always natural or obvious. Thus, re-representation is often a key step in the creative process.

Re-representation is also a vital part of teaching. Al-Khwarizmi explained algebra through complex geometric word problems, yet the modern representation of algebra as alphanumeric equations is simple enough to teach children.

We present *task distillation* as a computational model of re-representation designed for teaching. This generalization of dataset distillation (Wang et al. 2018) involves transforming a given learning task into a smaller, more quickly learned synthetic task that can be used to train a model such that the model's performance approximates the performance of a model trained directly on the original task. The synthetic task is a highly compressed representation that is more information-dense than the original task's representation. Most machine learning tasks rely on naturalistic data representations that sample from some real-world data distribution, while the synthetic task is free from naturalistic constraints and can be significantly reduced in size. To provide a concrete example of task distillation, we distill reinforcement learning environments into single-batch synthetic supervised learning datasets that can be learned in a single step of stochastic gradient descent (SGD). We show that the task distillation meta-learning process creates a new representation, the synthetic task, capable of being used to teach the task to a variety of learners. We do not claim that learners shown in our simple examples develop creative solutions during this process, but we argue that they do gain the task-specific expertise that is a precursor to potential creativity. In addition, this new representation is a compact cognitive artifact that can aid in understanding the original task. This small, information-dense representation may be easier to manipulate than the original representation in searching for creative solutions.

Re-Representation

Re-representation is a process by which the features of a given task or object are transformed from their direct representation into another form. The target representation is useful if it can be manipulated into creative solutions in a more obvious way than the original representation (Wiggins and Sanjekdar 2019). This is related to transformational creativity in Boden's theory of creativity: re-representation transforms the creative space of the original representation, such that more intuitive exploration for creative artifacts can be performed (Boden 1990).

Re-representation can be a purely internal process: transforming stimuli to match a representation in memory held in the brain. This is analogous to "seeing as", interpreting a novel stimulus as a familiar object, which can be mentally manipulated (Olteţeanu 2015). However, re-representation can be realized externally, by manipulating physical or meta-physical objects. For example, a sculptor must physically deform stone with a chisel to realize their internally-represented vision of the sculpture. While

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

this final physical representation is the creative artifact, not all re-representations must be creative artifacts; but even re-representation-as-pure-intrinsic-mental-state might be intentionally externalized as an artifact.

Such a re-representation of a task can be utilised to teach basic competency in the task and/or to facilitate further creative problem solving; and this re-representation can be in an entirely different domain than the task itself. The Atari game *Centipede*, for example, requires no natural language skills (Atari 1980), yet expertise in *Centipede* can be taught primarily through natural language. *The Video Master's Guide to Centipede* is one such example, with over 100 pages of natural language and diagrams outlining complex strategies for maximizing score. The guide provides such teaching without a single screenshot of the actual game (Dubren 1982). The author re-represented expertise in playing *Centipede* into natural language, and the reader must re-represent the language back into *Centipede* gameplay.

While teaching expertise can be a creative task in itself, expertise is required for any type of creative task. Expertise can be vital to intentional and efficient searches of a creative space, but it is even more necessary in proving an artifact is creative. While novelty and value are core determiners of creativity, it is the field of the domain which judges novelty and value. Without being able to demonstrate expertise to the field, an otherwise creative artifact will have no impact on the field and be forgotten (Csikszentmihalyi 1996). Expertise can be held by an individual or be distributed throughout a system, but creativity cannot occur without expertise (Reilly 2008).

Task Distillation

In task distillation, one learning task is re-represented as a separate synthetic task that can be used to teach expertise quicker than direct learning on the original task. This is a generalization of dataset distillation, extended to allow for machine learning tasks beyond supervised learning datasets (Wang et al. 2018). In order to demonstrate task distillation, we distill the cart-pole and Atari *Centipede* environments into single-batch supervised datasets. We provide a brief formal definition of task distillation, and then provide examples of its ability to teach through re-representation.

Task distillation consists of producing a synthetic task T_d from a target task T_0, such that T_d contains the compressed teaching potential of T_0 for a distribution of learner models. That is, learners trained on T_d should approximate the performance on an evaluation task of learners trained directly on T_0. In addition, the distilled task should be a compressed representation of the original task: $|T_d| \ll |T_0|$. Thus, T_d can be used instead of T_0 for training to reduce training costs without a significant drop in performance.

Task distillation is not limited to simply compressing a task into a denser representation of the knowledge required to teach. Rather, it can be used to transform a learning task into a different modality. We demonstrate one form of trans-modal task distillation by distilling reinforcement learning (RL) environments into synthetic supervised learning (SL) datasets. As a consequence of re-representing an RL environment as an SL dataset, new learners will not need to explore an environment to learn the task once the distilled dataset is created. In our examples, the learners can achieve expertise on the original task by training on the distilled dataset in a single step of stochastic gradient descent with mean squared error loss—significantly cheaper than the reinforcement learning process it replaces.

Methods

We provide experiments that distill two environments: cart-pole and Atari *Centipede*. First, we informally describe our algorithm for generalized task distillation, though other dataset distillation algorithms could also be generalized to this end. Second, we provide implementation details used in our experiments.

Our method for task distillation is based on the meta-learning method for dataset distillation (Wang et al. 2018). This method utilizes a nested loop: the inner loop trains a new learner on the distilled task, and the outer loop uses the trained learners' performance on the real task to update the distiller to teach the learners to better perform on the task. We provide a diagram (Figure 1) to show the meta-learning process for distilling an arbitrary task into a synthetic task. A formalization of the algorithms for task distillation and RL-to-SL distillation is beyond the scope of this work.

In each experiment, we distill a reinforcement learning environment for a set of learners with the same architecture. We utilize proximal policy optimization (PPO) as the outer loss function, and include an auxiliary critic network that is optimized alongside the distiller on the PPO loss. The critic is required for calculating PPO policy loss and is discarded when training is completed (Schulman et al. 2017). The architectures and hyperparameters are standard for direct-learning PPO on cart-pole and Atari, respectively. [1]

We create our distiller by parameterizing a randomly-initialized dataset of the dimensions we want for our final dataset. Each instance must match the dimensions of the target environment's state space in order to fit in the learner networks. The number of instances in the final distilled dataset is a hyperparameter and can only be optimized through experimentation. The synthetic data instances are updated directly by the optimization algorithm. Soft labels are used and optimized (Sucholutsky and Schonlau 2021b), being represented as a vector matching the size of the environment's action space. Other formulations are possible, such as the generative teaching network (GTN). This formulation involves training a generator network to produce the distilled data, allowing for more than one dataset to be produced. The GTN is capable of distilling cart-pole (Such et al. 2020); however, training the distilled set directly appears to be more effective, especially for Atari environments. In addition, we preferred a single high-quality re-representation, while the GTN generates many varied re-representations.

For cart-pole, distillation success was determined by

[1] See the following blog for standard PPO implementation details: `https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/`. We utilized all these details except for learning rate annealing and value loss clipping for both our direct RL and distillation experiments.

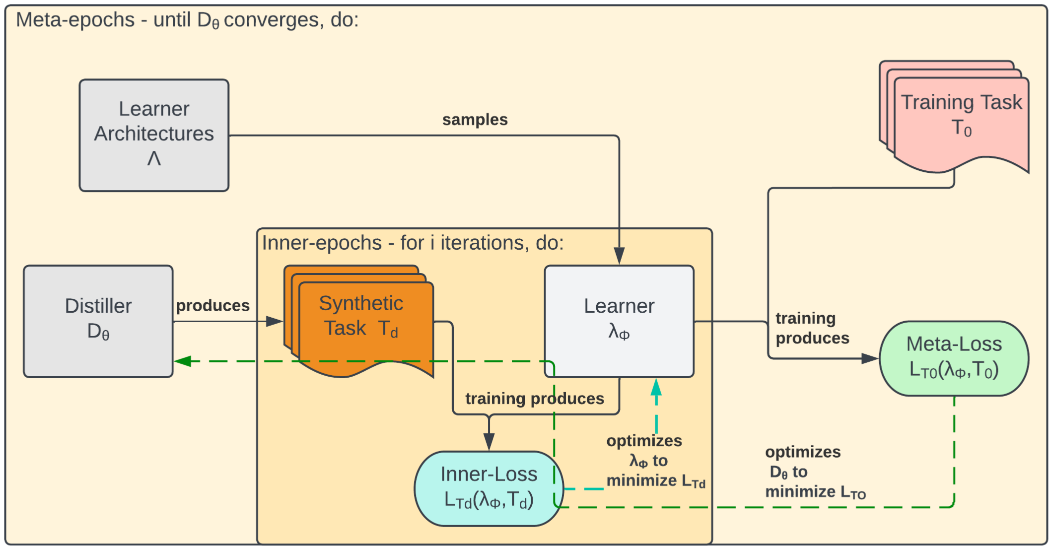


Figure 1: The meta-learning process for generalized task distillation. The inner loop involves training a newly sampled learner on the synthetic task. The trained learner is tested against the real task, and the loss is backpropagated through the inner learning process back into the distiller. This repeats until the distiller converges, which can be seen by the learners' average performance.

whether the distilled data could teach randomly sampled models from the learner set to fully solve the task (defined somewhat arbitrarily as attaining a reward of 500). For *Centipede*, which does not have a "solved" state (but rather is a more open-ended problem of score maximization), we compare the average reward reached by distiller-trained models to the reward reached by a PPO agent. This is reasonable because distillation's meta-learning relies on the same loss function as direct learning and thus has the same limitations.

Cart-pole Distillation

As a standard toy problem for reinforcement learning, cart-pole demonstrates the advantages of distillation and shows how distillation re-represents the entire environment as a small dataset that can be used to understand the cart-pole task. The cart-pole problem is a simple classical control problem based on a physical system. A cart is connected to a pole by a hinge. The pole begins nearly perfectly upright. The agent must move the cart either left or right each timestep to attempt to keep the pole balanced atop the cart. If the pole rotates past a certain threshold in either direction, or if the cart moves past a threshold, the simulation ends. The goal is to balance the pole as long as possible, up to a maximum of 500 timesteps.

This task is easily solved by deep reinforcement learning. We distill this problem into a single-batch representation for supervised learning using randomly initialized agents with the same architecture. The resulting distillation set can be used to train all models sampled from the learning set to balance the pole up to the time limit, solving cart-pole.

It took approximately 3.5 times the number of cart-pole episodes for the distillation to converge compared to direct RL learning. With the increased overhead of meta-learning, distillation was approximately 6 times slower. However, the end result of distillation is a single 2-instance dataset that can teach the cart-pole task in one SGD step (see Figure 2 for a visualization). Thus, distillation can be a cheaper alternative to sequentially training 6 or more RL agents. In addition, the distilled dataset is an artifact that can be used to more easily interpret the original task, simplifying cart-pole's infinite state space into two key examples.

We have experimented with a variety of distilled dataset sizes and have determined that all dataset sizes, above a certain threshold, are capable of being used to solve the problem. The minimum sized teaching set is most interesting, as it is the densest learning representation possible using distillation. Interestingly, this also appears to be more human-interpretable, providing an explainability artifact that shows the "essence" the task. As shown in Figure 2, cart-pole's continuous state space is distilled into two discrete examples that completely characterize the task: showing the pole leaning left in one and right in the other. Neither the state transition function nor the reward function are directly modeled; instead, the action labels clearly demonstrate that to maximize reward the cart must simply be moved in the direction the pole is leaning. While this does not explain the whole model of the system's physics, it shows how to move the cart to balance the pole, which is all that is needed to solve the task. In addition to its explainability potential, the 2-example distilled dataset is also the cheapest learning representation, though negligibly so compared to other one-batch distilled datasets.

For cart-pole, this minimum teaching dataset contains only two instances. This is the theoretical limit for environ-

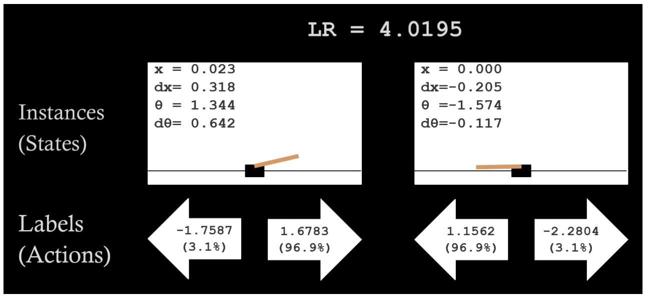


Figure 2: The minimum-sized distilled teaching set for the cart-pole environment. Training on this set for a single step of SGD can teach the cart-pole task to any member of the learner set. The state vectors are shown numerically and visually. The action labels are provided as raw values as well as a softmaxed policy for the provided state. Note that the state is not a valid cart-pole state: the environment would have ended after the pole reached $\theta = \pm 0.2095$, and the simulator does not work for values beyond $\theta = \pm 0.418$. This demonstrates that the distilled instances are not copies of data seen during distillation training; they are synthesized.

ments with only two actions that are required for solving the problem; the teaching set must provide a distinction between when to use these two actions (Sucholutsky and Schonlau 2021a). Notice that in the continuous state space of cart-pole, there are virtually infinite possible states, restricted only by the computer's precision. However, as we can see in Figure 2, the strategy for cart-pole can be described in two states: the cart should move left with a left-leaning pole, and right with a right-leaning pole. While this simple strategy does not address edge-cases, such as when the cart is near the edge of the screen, it is still sufficient to solve the problem.

Centipede Distillation

The Atari 2600 environments represent a significant increase in difficulty from cart-pole by greatly increasing the state space dimensionality, the action space, and the complexity of strategies required to perform well on the environment. We demonstrate that complex reinforcement learning environments can be distilled by successfully distilling a teaching dataset of only 10 instances, which can be used to train the learners to perform well on *Centipede*. This is the theoretical minimum sized dataset required to teach *Centipede*; given we are using soft-label vectors and *Centipede* has 18 distinct actions (Sucholutsky and Schonlau 2021a).

Unlike cart-pole, there is no well-defined solution to *Centipede*: a player's goal is to maximize score. Reaching the theoretical maximum score is well beyond the capabilities of small reinforcement learning agents, given our resources. Therefore, we judge distillation success by comparing an individual's cumulative reward on *Centipede* after training on the distilled task versus training on *Centipede* directly.

Direct learning on *Centipede* yields an average reward of 9167 points after approximately $1,000$ epochs of training. Distillation yields an average reward of 8083 points on *Centipede* after approximately $8,000$ epochs of training, reaching 88% of the reward in 8 times the number of optimization steps. The drop in average reward is expected: we are compressing knowledge gained from testing against *Centipede*; it is unlikely that a distillation of a task can be used to teach a learner to perform at a higher level than can the original task itself. However, the learners still perform well above random, and the best-performing learner trained on that distilled data achieved a score of $36,978$—well above the human average of $11,963$ (Mnih et al. 2015).

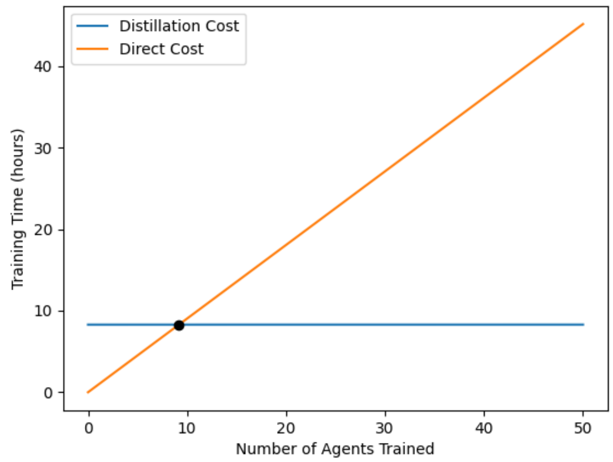


Figure 3: Time costs for training *Centipede* agents using distillation versus direct task learning. While training the distiller is costly, distillation training time increases negligibly (by 0.18 seconds) as the number of agents trained increases. Direct learning is significantly cheaper for a single agent but must be repeated in full for each additional agent; when training more than 9 agents, distillation is cheaper. As the number of agents trained increases, distillation becomes more cost-effective compared to direct learning.

The cost increase is also expected: distillation pays much of the learning cost up-front. Each epoch of direct RL training on *Centipede* using our resources took on average 3.25 seconds, compared to an average of 3.73 seconds per epoch for distillation. While the distillation process takes approximately 9.18 times as long as direct RL training ($8,000$ epochs x 3.73 seconds/epoch vs $1,000$ epochs x 3.25 seconds/epochs), the benefits of training on the distilled set is clear. Training on the distilled data is significantly cheaper than the full direct RL training. It takes on average 0.18 seconds to train a model on the distilled data, 18,000 times faster than RL training. This speedup is due to distillation removing the requirement to interact with the environment, as well as the amount of data trained on: 10 instances for training on the distilled data versus 8,000,000 instances for training on the environment. Using our resources, it is more time-effective to utilize distillation rather than directly learning on *Centipede* if one is training more than 9 models. See Figure 3 for the training time costs of distillation and direct learning on *Centipede* as a function of number of learners trained.

Similar to the distilled cart-pole instances, the distilled *Centipede* instances represent invalid states. Their values go beyond the range of valid pixel values and cannot be accurately represented as images. However, despite not being intuitive representations that resemble real *Centipede* states, these representations are capable of teaching the learners. While this representation is not as easily interpretable as the cart-pole distillation, it re-represents the task to efficiently impart expertise to the learners. Even so, the dataset provides another artifact that can be examined alongside the environment and the agents to more fully explain the learning process on the environment and potentially lead to creative behavior invention.

Discussion

Our experiments demonstrate the re-representation of the cart-pole and *Centipede* learning tasks as compressed representations that teach through a different learning mode. The re-representations do not contain all information about the original environments: there is no indication of the range of states, the state-transition function, or the reward function. Rather, only pertinent learning information is stored.

While the systems described in this work are not creative, the systems contain the expertise which is a necessary prerequisite for creativity (Reilly 2008). The expertise, gained through many iterations of learner training and testing, is aggregated within the distilled task. Upon learning, the expertise is imparted to a learner which can perform the targeted task. If one needed expertise in a creative system, a pre-distilled set is a much quicker alternative to gaining expertise by learning on the whole task. With a dataset distilled from a reinforcement learning environment, a model can be trained in seconds rather than hours of exploration. The resources required to explore the environment's state space to gain expertise can instead be used toward exploring a creative space. In addition, this re-representation provides another way to understand the environment, one which could be manipulated to allow for the invention of creative and interesting strategies in the environment.

For example, consider a simple creative system that utilizes distillation to create a cart-pole agent capable of performing tricks, which could be used in a novel balancing routine. Without distillation, this might be done by providing a variable reward function, which is changed to reinforce policies that lead to interesting and novel behavior, as judged by a separate evaluation function. The space of reward functions can then be searched to find reward functions that result in producing creative behaviors (as judged by the evaluation function). However, without distillation, each point in reward function space can only be tested by a full session of (expensive) reinforcement learning. Distillation can be used instead—the search can be performed directly on the distilled training set's parameter space. Testing a point in this re-represented space can be performed more efficiently than using RL: one inexpensive SGD step and one episode of performance on cart-pole to demonstrate the learned behavior and receive a score from the evaluator model.

A search for creative and interesting strategies in *Centipede* can benefit from distillation in much the same way as cart-pole. Searching through the smaller distillation space, compared to the parameter space of a complex reward function, as well as cheaper training for evaluation, would provide a significant speedup to the creative search. The time saved evaluating each point could then be put toward searching more points in the space, allowing for a more thorough examination of the creative space, and potentially finding a superior creative artifact than could be found using the same resources without distillation.

References

Atari. 1980. Centipede. Arcade, Atari 2600.

Boden, M. 1990. *The Creative Mind*. Abacus.

Csikszentmihalyi, M. 1996. *Creativity*. HarperCollins Publishers, 1st edition. chapter 2, 27–31.

Dasgupta, S.; Papadimitriou, C.; and Vazirani, U. 2006. *Algorithms*. McGraw-Hill Education. chapter 0, 2.

Dubren, R. 1982. *The Video Master's Guide to Centipede*. Bantam Books.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature* 518:529–533.

Olteţeanu, A.-M. 2015. "Seeing as" and re-representation: Their relation to insight, creative problem-solving and types of creativity. *Publications of the Institute of Cognitive Science*.

Reilly, R. C. 2008. Is expertise a necessary precondition for creativity?: A case of four novice learning group facilitators. *Thinking Skills and Creativity* 3(1):59–76.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv* 1707.06347.

Such, F. P.; Rawal, A.; Lehman, J.; Stanley, K. O.; and Clune, J. 2020. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *Proceedings of the 37th International Conference on Machine Learning*, 9206–9216.

Sucholutsky, I., and Schonlau, M. 2021a. "Less than one"-shot learning: Learning N classes from M < N samples. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, 9739–9746.

Sucholutsky, I., and Schonlau, M. 2021b. Soft-label dataset distillation and text dataset distillation. In *Proceedings of the International Joint Conference on Neural Networks*, 2220–2227.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. 2018. Dataset distillation. *arXiv* 1811.10959.

Wiggins, G. A., and Sanjekdar, A. 2019. Learning and consolidation as re-representation: Revising the meaning of memory. *Frontiers in Psychology* 10.

Transformational Creativity Through the Lens of Quality-Diversity

Jonathan Demke[1], Kazjon Grace[2], Francisco Ibarrola[2] and Dan Ventura[1]
[1]Computer Science Department, Brigham Young University, Provo, UT 84602 USA
jpd0057@byu.edu, ventura@cs.byu.edu
[2]School of Architecture, Design and Planning, The University of Sydney, Sydney, Australia
kazjon.grace@sydney.edu.au, francisco.ibarrola@sydney.edu.au

Abstract

Quality-Diversity algorithms are a useful tool for creative search because they can evolve artefacts with different search strategies that focus on different criteria such as behavioral novelty or improving quality. These different strategies find diverse solutions with high quality. However, Quality-Diversity algorithms can exhibit inefficiency and a lack of control when exploring a genome space for diverse solutions. We propose two different approaches that connect the genome space of artefacts to their phenotypic behavior. The approaches may allow a computationally creative system not only to explore the space of solutions more quickly but also to guide its search towards underrepresented and surprising behaviors.

Introduction

Artefact-producing computational creative (CC) systems focus on generating artefacts that exhibit qualities like typicality, novelty, and quality. Developing metrics to evaluate these criteria is a challenging task, and so is finding the right balance for this multi-objective optimization problem: an insufficient focus on novelty would lead to artefacts that are merely good; an insufficient focus on typicality would lead to artefacts that are merely weird; an insufficient focus on quality would lead to artefacts that are merely artefacts. Accordingly, the subject of creative search has long been a focus of CC research (Boden 1992; Wiggins 2006). A key concept in this debate has been the contrast between exploratory and transformational creativity, with the former involving the search of a pre-defined space, and the latter being search that combines both redefinition and search of a space. Transformational creativity, viewed as the more significant of the two, is motivated by studies in the cognitive science of creativity showing iterative problem re-framing and the gradual emergence of both a problem and its solution (Maher and Poon 1996; Schon and Wiggins 1992). The basic nature of search algorithms implies a fixed space, and creative search has been operationalized as a kind of "meta-search"—a higher-level search over the space of search spaces permitting some level of transformational creativity in the original space.

Quality-Diversity (QD) algorithms are evolutionary algorithms (EA) that simultaneously optimize one fitness function (for "quality") while producing a set of solutions with good coverage of one or more other functions (or "behaviors", for "diversity"). QD algorithms only allow a solution to compete against its neighboring solutions; for example, it may not make sense to compare Edgar Allen Poe's horror poems to Shel Silverstein's comedic poems, but we may aptly compare Poe's poems with Emily Dickinson's horror poems. They enforce diversity by developing behavioral niches to stratify solutions while also finding quality solutions within those niches, leading to a collection of diverse solutions, each with high quality relative to their neighbors. The QD approach is not strictly a kind of meta-search, having a static genome space and being better understood as a kind of multi-objective search for a set of solutions. However, we argue that it exhibits some of the same properties, in that the diverse high-quality artefacts it tends to produce were evolved using different search strategies (i.e. combinations of the behavior and fitness functions), and tend to achieve their quality in very different ways (as the behavior space is likely nonlinear with respect to the genome space). QD algorithms are particularly useful in a co-creative CC context, where a set of diverse, high-quality options can be presented to a user, resulting in increased novelty in the final product and offering the potential for transformational creativity from the perspective of the user, who may have preconceptions about the space being searched.

Unfortunately, QD algorithms often struggle to search complex spaces thoroughly in reasonable amounts of time, as they make the assumption that the fitness and behavior evaluations can be run millions of times (Chatzilygeroudis et al. 2021). One approach to addressing this is introducing a surrogate evaluation—some rapid approximator of the more-expensive fitness and/or behavior functions (Ong, Nair, and Keane 2003). While surrogates have been applied to some CC domains (Zhang et al. 2022), for many domains the evaluation functions—as well as the expression from genotype to phenotype—are both expensive and challenging to approximate. As the dimensionality of the genome and the number of desired behaviors increases, many QD algorithms expend considerable resources exploring and re-exploring well-travelled and low-potential regions.

In this paper we propose two different approaches to improve QD search, particularly motivated by the CC context. Both approaches are based on attempting to predict how the

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

behavior space will change based on changes in the genome space. Our first approach is a low-level and local one: learning local gradients of individual behaviors and adapting individuals along them. Our second approach is a high-level and global one: trying to discover latent behavioral structure in the genome space.

Background

MAP-Elites (Mouret and Clune 2015) is one of the most influential QD algorithms to date. MAP-Elites maps an individual to a point in the behavior space $\mathbb{B}$, which specifies different behavioral qualities of the individual's phenotype. For example, we could map the space of stories to k behaviors, such as the amount of humor, horror, and romance contained within the story. MAP-Elites then searches the genome space to find quality solutions within behavioral niches such as the best story with a lot of humor and romance, but little horror.

The approach requires definition of a genome space $\mathbb{G}$; phenome space $\mathbb{P}$; genome to phenome map $T : \mathbb{G} \to \mathbb{P}$; behavior function $B : \mathbb{G} \to \mathbb{R}^k$, where k is the number of behavioral attributes; fitness function $P : \mathbb{G} \to \mathbb{R}$; and an archive of the elite solutions $(\mathbf{P}, \mathbf{X})$, where $\mathbf{P}$ stores the fittest score for each behavioral niche and $\mathbf{X}$ stores the fittest individual for each behavioral niche.[1] For example, let $\mathbb{P}$ be the domain of stories, $\mathbb{G}$ some genomic representation (such as a plot graph), T a story generator, and $k = 3$ behaviors: humor, horror, and romance. If B maps each behavior to a real value between 0 (e.g. not scary) and 1 (e.g. the scariest), then our behavior space $\mathbb{B} = [0, 1]^3$. $\mathbb{B}$ is then discretized along all k dimensions to form the phenome's behavioral niches. If horror and romance are discretized into bins, $[0, 0.5), [0.5, 1]$, meaning a story is considered either not scary (not romantic) or scary (romantic), and humor is discretized into bins, $[0, 0.33), [0.33, 0.66), [0.66, 1]$, so that a story can be mapped to low, medium, and high levels of humor, the result is twelve behavioral niches.

MAP-Elites randomly samples an individual $g \sim \mathbb{G}$, retrieves its behavioral niche $b \leftarrow B(g)$ and performance $p \leftarrow P(g)$, and checks if g is the fittest within its behavioral niche $p > \mathbf{P}[b]$. If it is, then the archive of elites is updated: $\mathbf{P}[b] \leftarrow p$ and $\mathbf{X}[b] \leftarrow g$. After enough random samples, MAP-Elites starts searching the genome space by performing genetic operations, e.g. crossover and mutation, among the elite solutions in $\mathbf{X}$.

Approximating Behavior Gradients

When a genome is mutated, the resulting child g usually exhibits a change in behavior Δb. Unfortunately, it is difficult to determine what specific change Δg led to Δb. Furthermore, it is unclear whether additional mutation in the direction of Δg leads to additional behavior change in the direction of Δb. This is a credit assignment problem. Our local, "low-level" approach is an attempt to alleviate this problem by approximating the gradient of our behavior function, ∇B, so that we can correctly assign credit to each gene.

[1] B and P measure *phenotypic* behavior and fitness, respectively, so each includes an implicit use of the mapping T.

By utilizing a differentiable regression model $f_\theta : \mathbb{G} \to \mathbb{B}$ as a surrogate for B, we can approximate ∇B by instead computing ∇f_θ. To compute f_θ, a genome $g \in \mathbb{G}$ can be mutated to generate neighbors g_i; their corresponding behaviors b_i retrieved; and f_θ trained to minimize error, i.e. $\min_\theta \| f_\theta(g_i) - b_i \|$. We can then either apply $\nabla f_\theta(g)$ directly: $g \leftarrow g \pm \nabla f_\theta(g)$ or increase the mutation rate in the direction of $\nabla f_\theta(g)$. Biasing mutation in the direction of maximal expected behavioral change could reduce the time an algorithm like MAP-Elites (which applies Gaussian noise as its mutation operator) spends searching regions of the genome space that have little chance of producing improvements in either quality or diversity.

The utility of this approach as an efficiency improvement would depend on the amount of data required to train a local behavioral regressor and the size of the region in genome space that said regressor could reasonably approximate gradients over. If a single global f_θ is accurate enough to resemble B, then utilizing ∇f_θ over the entire genome space would significantly speed up search. This might be possible if an underlying structure between $\mathbb{G}$ and $\mathbb{B}$ exists; that structure could be discovered with a neural network f_θ. However, it is unlikely that a single global f_θ will suffice, and therefore it may be necessary to employ multiple local regression models to approximate B in piecewise fashion.

A naïve first approach would define some radius around a genome and build a regressor on the mutations taken around the genome. If the search ever moves beyond the radius of the genome then we create a new regressor. Simple linear regressors could be trained with few data examples to give a quick approximation of a genome's local behavior gradient.

It might also be useful to utilize both global and local regressive models. Ensemble disagreement (Lakshminarayanan, Pritzel, and Blundell 2017) or randomized prior functions (Osband, Aslanides, and Cassirer 2018) can be built with regressive neural networks to simulate Bayesian uncertainty, which can allow EAs to exploit the global regressive model's gradient approximation when the global model's uncertainty is low and utilize a local regressive model when its uncertainty is high.

Our approach is comparable to natural evolution strategies (Wierstra et al. 2008) and covariance matrix adaptation evolution strategies (Hansen 2016), which also adapt mutation towards an approximate gradient; however, by uncoupling our gradient approximators from the current search we can use them for purposes other than finding the next population artefacts, such as backtracking or exploring the genome space on a different behavioral axis. Our approach also includes the possibility of using a global gradient approximator. Local gradient approximation is also analogous to "local explanation", an approach used as a form of explainable ML such as LIME (Ribeiro, Singh, and Guestrin 2016), which model the local environment around a datapoint using a simple, scrutable model, allowing the reasons for its classification to be made clear.

Even with gradient approximation, it can still be difficult to navigate the genome space to find some expected behavior—sometimes moving towards one behavior axis can move you away from another behavior axis.

Learning a Genome-Behaviour Latent Space

There may exist behaviorally-induced global latent structure within the genome space that may be discoverable during search. For example, a variational autoencoder (VAE) could be utilized to construct a latent space that is easier to explore (than genome space), because it clusters the genomes behaviorally, allowing sampling from the VAE's simple prior distribution to get genomes within each cluster—sampling and decoding from the latent space facilitates "intelligently" jumping around the genome space. To ensure the VAE's latent structure captures the desired behavior, the prior, encoder and decoder may be conditioned on that behavior (Sohn, Yan, and Lee 2015).

There are two challenges we see in this approach. First, VAEs commonly use continuous latent distributions to represent the data, most notably the multivariate Gaussian, which typically has a smoothness artefact that biases the mapping from similar latent values toward similar decoded outputs; however, dissimilar genomes may share behavioral features. They can also suffer from posterior collapse, where an overparameterized decoder will largely ignore most of the latent structure. In such cases, discretized latents, such as vector-quantized or categorical latents, may prove useful for alleviating these issues (van den Oord, Vinyals, and Kavukcuoglu 2017; Hafner et al. 2021).

The second and potentially more challenging issue is retrieving the necessary data to train the VAE; VAEs learn by maximizing the evidence lower bound, but what serves as the evidence for the genome space? If all genomes are considered equally likely, then the expected fitness of a VAE sample should approximate the expected fitness of the entire genome space, which may be extremely low in large genome spaces. The VAE likely wouldn't be a useful tool in this scenario. A possible solution to this problem might be to weight genomes based on their fitness, similar to how EAs perform parent selection; however, careful attention is required to ensure that the few high-performing genomes do not dominate as the evidence, since the VAE will likely overfit on the few samples and not generalize to other high-performers in the genome space. Similarly, we could weight genomes in underrepresented behavioral niches more heavily to enhance the coverage of behaviors.

Assuming a well-trained VAE model, with likelihood (encoder) $p_\theta(z \mid g, b)$, prior $p(z \mid b)$, and posterior (decoder) $q_\phi(g \mid z, b)$ distributions, one can sample from the model to find elites in the neighborhood of other elites: $g' \sim q_\phi(g \mid z, b)$ where $z \sim p_\theta(z \mid g, b)$ is a sample near the encoding of an elite $g \in \mathbf{X}$ with corresponding behavior niche b. We can also sample for a new elite in a behavioral niche b where no elite exists yet, by sampling our prior $z \sim p(z \mid b)$ and retrieving a genome from our posterior $g \sim q_\phi(g \mid z, b)$. It is important to note that although $g \sim q_\phi(g \mid z, b)$ is conditioned on behavior b, it does not guarantee that $B(g) = b$. However, this allows us to understand where behavior is not well understood within our models; we can use VAE samples to analyze whether their true behavior matches the given conditional behavior as a way to measure the information gap of behaviors in the genome space, e.g. $\mathbb{E}_{z \sim p(z|b)}\left[\|b - B(q_\phi(g \mid z, b))\|\right]$.

Genome-Behavior Models and QD Curiosity

In complex genome spaces, the number of datapoints required to train the models in either approach may be large enough to eliminate any efficiency gains. Yet if either of these approaches is effective at connecting the genome and behavior spaces in QD applications—and we stress that *if,* because at present we haven't tested either approach beyond toy problems—then there may be more benefit to CC than any gains in efficiency. Maximizing coverage of one or more behavior functions is interesting, in that it offers a stepping stone to more CC-relevant concepts like novelty, but outside of the co-creative "offering diverse suggestions to a human" use case it is actually somewhat conceptually unsatisfying as a step towards creative search.

Creative search—and the transformational creativity it seeks to enable—are fundamentally motivated by the discovery of specific radically new solutions. The constant outward pressure of QD algorithms, however, values the entire behavior space equally at all times. Radically new solutions may emerge but are treated no differently than incrementally more-fit or more-diverse ones. This evokes 1960s "ideational fluency" notions of creativity (Torrance 1966), in the sense that QD algorithms produce the largest possible set of meaningfully different solutions to a problem. Being only a half-century behind the psychologists is still not bad for a CC algorithm, but it's possible we can do better by rethinking the problem definition for a CC context.

In classic QD algorithms the selection of where to search next is random—either by selection of a random existing elite in MAP-Elites and its differentiable derivatives like OMG-MEGA (Fontaine and Nikolaidis 2021), or by sampling from a distribution learned over the genome space in evolutionary strategy derived QD approaches like CMA-ME (Fontaine et al. 2020). This randomness seems unavoidable: QD algorithms are driven to explore the behavior space but cannot act directly within it—they must instead search the genome space and hope that doing so illuminates new behavior. The approaches we propose in this paper, however, offer an opportunity for a curiosity-motivated QD algorithm, grounded in a connection between genome and behavior. A "Curious Quality Diversity" algorithm might choose where to search next (within the genome space) based not on direct predictions of behavior, but on a drive to improve the quality of those predictions. This curiosity drive could be used to dynamically nudge QD search towards regions where behavior is not well understood—which may indicate potential for radically new (and potentially high-quality) artefacts.

The term "curiosity" has been used in QD algorithm selection before, with the "curiosity score" assigned to each elite in (Cully and Demiris 2017) being the expected probability that selecting that elite and mutating it would lead to offspring that are themselves elites (i.e. are either sufficiently different to all known elites or better than all elites they are similar to). In a sense, this is a model of what Berlyne would call "general curiosity", the drive towards any new stimulus (Berlyne 1960), which is consistent with the overall aim of QD algorithms. By contrast, "Curious QD" gives preference to new individuals that would improve the system's model of the behavior space, consistent with Berlyne's "specific curiosity" and other similar "learning progress" notions (Oudeyer 2004; Schmidhuber 2010; Grace and Maher 2015).

While we admit to not having yet implemented any of these ideas, "Curious QD" could be implemented using either of our above approaches for connecting behavior and genome space combined with techniques from the field of Bayesian optimization (BO). BO techniques are active learning approaches that (when applied to learning a Bayesian ML model like a Gaussian Process) define an information-theoretic acquisition function over where to look next. Typical acquisition functions include upper confidence bounds (i.e. picking the spot that could theoretically be best, given uncertainty) and expected information gain (i.e. picking the spot that will reduce uncertainty the most). Applied to either of our proposed approaches, which would by necessity be learned in an active learning context, these BO techniques could produce the kind of medium-term search dynamics more recognizable as specific curiosity.

Conclusion

We have proposed two approaches that work in conjunction with the QD algorithm MAP-Elites. Our approaches focus on connecting a genome space to its phenotypic behavior space, either by approximating the local gradients of the behavior functions, or by finding a latent structure that correlates the genome space with the behavior space. These approaches not only may promote efficiency of creative search, but also, by modelling the behavior space, they offer an ability to control *how* a CC system explores that space. We also offer some initial thoughts about "Curious QD" and how a CC system could utilize these models of the behavior space to find specific, radically novel artefacts with high-quality. Although this work is still in the preliminary stages, it appears encouraging as a way to think about and operationalize the concept of transformational creativity.

References

Berlyne, D. E. 1960. *Conflict, Arousal, and Curiosity*. McGraw-Hill Book Company.

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Chatzilygeroudis, K.; Cully, A.; Vassiliades, V.; and Mouret, J.-B. 2021. Quality-diversity optimization: A novel branch of stochastic optimization. In *Black Box Optimization, Machine Learning, and No-Free Lunch Theorems*. 109–135.

Cully, A., and Demiris, Y. 2017. Quality and diversity optimization: A unifying modular framework. *IEEE Transactions on Evolutionary Computation* 22(2):245–259.

Fontaine, M., and Nikolaidis, S. 2021. Differentiable quality diversity. In *Advances in Neural Information Processing Systems*, 10040–10052.

Fontaine, M. C.; Togelius, J.; Nikolaidis, S.; and Hoover, A. K. 2020. Covariance matrix adaptation for the rapid illumination of behavior space. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 94–102.

Grace, K., and Maher, M. L. 2015. Specific curiosity as a cause and consequence of transformational creativity. In *Proceedings of the International Conference on Computational Creativity*, 260–267.

Hafner, D.; Lillicrap, T. P.; Norouzi, M.; and Ba, J. 2021. Mastering Atari with discrete world models. In *Proceedings of the International Conference on Learning Representation*.

Hansen, N. 2016. The CMA evolution strategy: A tutorial. *ArXiv* abs/1604.00772.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 6405–6416.

Maher, M. L., and Poon, J. 1996. Modeling design exploration as co-evolution. *Computer-Aided Civil and Infrastructure Engineering* 11(3):195–209.

Mouret, J.-B., and Clune, J. 2015. Illuminating search spaces by mapping elites. *ArXiv* abs/1504.04909.

Ong, Y. S.; Nair, P. B.; and Keane, A. J. 2003. Evolutionary optimization of computationally expensive problems via surrogate modeling. *AIAA Journal* 41(4):687–696.

Osband, I.; Aslanides, J.; and Cassirer, A. 2018. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 8626–8638.

Oudeyer, P.-Y. 2004. Intelligent adaptive curiosity: A source of self-development. In *Proceedings of the Fourth International Workshop on Epigenetic Robotics*, 127–130.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Schmidhuber, J. 2010. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* 2(3):230–247.

Schon, D. A., and Wiggins, G. 1992. Kinds of seeing and their functions in designing. *Design Studies* 13(2):135–156.

Sohn, K.; Yan, X.; and Lee, H. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 3483–3491.

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and Psychological Measurement*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 6309–6318.

Wierstra, D.; Schaul, T.; Peters, J.; and Schmidhuber, J. 2008. Natural evolution strategies. In *Proceedings of the IEEE Congress on Evolutionary Computation*, 3381–3387.

Wiggins, G. A. 2006. Searching for computational creativity. *New Generation Computing* 24:209–222.

Zhang, Y.; Fontaine, M. C.; Hoover, A. K.; and Nikolaidis, S. 2022. Deep surrogate assisted MAP-elites for automated hearthstone deckbuilding. In *Proceedings of the Genetic and Evolutionary Computation Conference*, 158–167.

Prompt diversification for iterating with text-to-image models

Francisco Ibarrola and **Kazjon Grace**
School of Architecture, Design and Planning
The University of Sydney
Sydney, Australia
[francisco.ibarrola,kazjon.grace]@sydney.edu.au

Abstract

The recent appearance of new generative models has transformed Creative Computing, allowing for the development of striking and original art and design. Nevertheless, achieving creative objectives depends heavily on supplying particular prompts for guiding the generation process. In this work, we use semantic models and affect to develop two methods to help the prompt building process, promoting exploration and subsequent specificity. We show some results obtained with these proposals and discuss the implications to image generation.

Introduction

Creative AI has been revolutionised by the recent emergence of new generative models that can produce visually stunning works of art and design (Rombach et al. 2022; Saharia et al. 2022) from a simple text prompt. However, this process is in practice rarely one-shot, as users iteratively refine their prompt, both to communicate a specific desired outcome to the model as well as (perhaps more importantly given what we know about the creative process) to refine and explore what it is they are after (Liu and Chilton 2021). In creative settings users often struggle to articulate their vision in precise enough terms, obtaining suboptimal results from generative models. This suggests an avenue for a new kind of co-creative interaction: suggesting prompt modifications to aid in this iterative exploration process. In this paper, we propose two novel approaches to address these challenges, intended to enhance the capabilities of creators working with generative models.

The first approach has to do with helping users refine their prompts to more accurately reflect their creative intent. The idea is based on Affect modelling (Osgood et al. 1975), a psychometrically validated approach which establishes three affective dimensions (Valence, Arousal and Dominance), quantifying peoples' feelings about a wide range of stimuli, including both words and images. This can be used to provide users with refinement suggestions that are diverse in terms of affect, and hence convey different impressions, guiding their creative process more accurately. Tapping into how words "feel" as opposed to (or in addition to) their semantic meaning provides an additional vector for prompt diversification.

The second approach is image-based, allowing users to provide a second image possessing certain attribute that they desire to imbue into their generated image but cannot quite grasp the term for. By identifying these underlying key points using image semantic latents (Radford et al. 2021) and presenting them as options, we enable users to guide the generative process towards their intent more precisely. Both of our approaches allow for greater creative control over the behaviour of generative models, but are also tuned towards generating more-diverse images and increasing the potential for serendipitous discoveries and creative pivots.

In the next section, we develop these two approaches in detail, and then provide some practical examples.

Prompt Modification Suggestions

Specificity enhancement

Let us consider a text prompt $\bar{y} \in \mathbb{Y}$ provided by the user as a first draft, on which we want to improve by suggesting some additional characterisation. Additionally, let us consider a set of words $Y = \{y_1, \ldots, y_N\} \subset \mathbb{Y}$ that are semantically similar to $\bar{y}$. This set can be constructed by means of the CLIP (Radford et al. 2021) encoder, which is a function $g : \mathbb{Y} \to \mathbb{R}^D$ that maps text prompts into a latent space, where similar vectors account for semantic similarity. In other words, Y can simply be built from a list of words maximising the normalized inner product

$$\langle g(y), g(\bar{y}) \rangle. \tag{1}$$

Given that all the elements in Y are close to $\bar{y}$ in the sense of (1), by transitivity they are all close to each other, and hence have some degree of semantic similarity. In order to propose meaningfully different suggestions to a user, we propose picking a subset of Y whose elements have different affect expressions.

To do this, let us consider the set of affect scores of these words, $A \doteq \{a_1, \ldots, a_N\}$, where $a : \mathbb{Y} \to \mathbb{R}^3$ is a function mapping a word to its three-dimensional affect score, and $a_n = a(y_n)$. Note that we are assuming we know the affect scores of these words, which can be extracted from a dataset or estimated using an approach such as the one proposed in (Ibarrola, Lulham, and Grace 2023).

We want to extract the "most diverse" subset $\hat{A} \subset A$ of K words to propose as enhancement options to the user. This

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

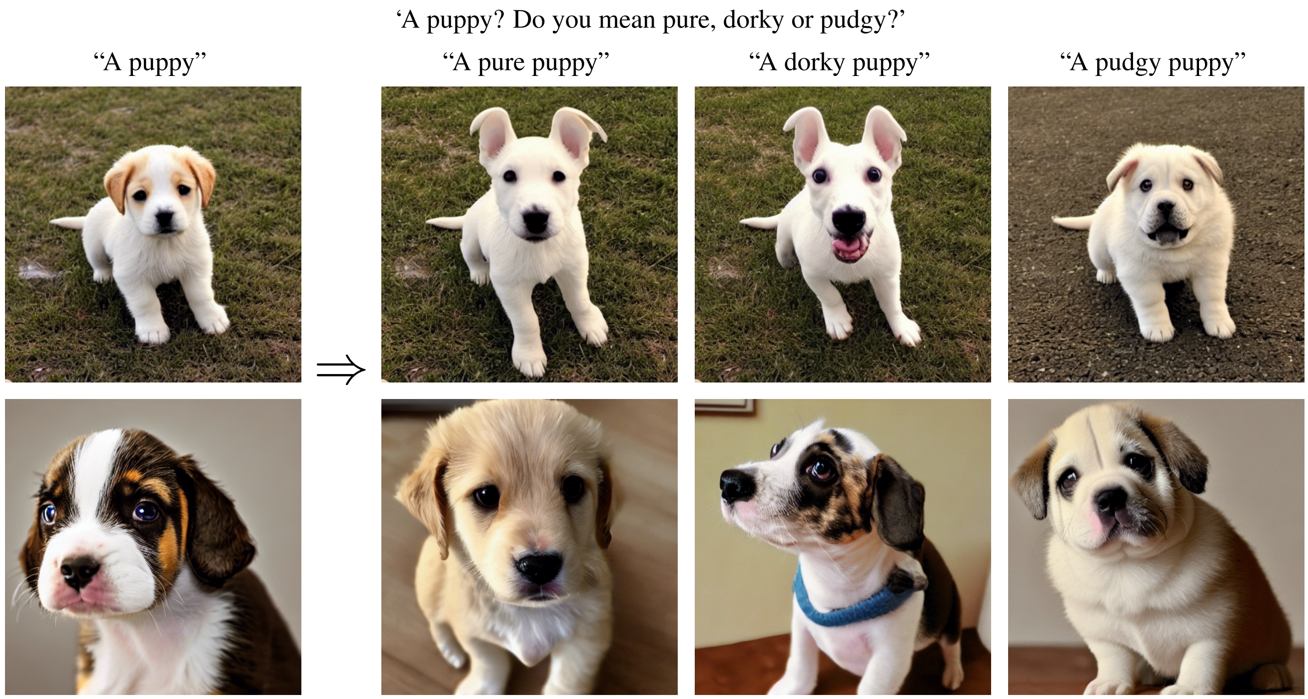


Figure 1: Illustration of two images generated by Stable Diffusion from the prompt “A puppy” (on the left), and those obtained after incorporating the enhancement suggestions made by our first approach (which in this case suggests “pure”, “dorky”, and “pudgy” as possible modifications. Each row was generated from the same random seed.

notion of diversity can be defined in many ways, depending how we choose to quantify it, and in this case we choose the largest minimum distance between the elements of a set. That is

$$\hat{A} \doteq \arg\max_{a_1,\ldots,a_K}(\min_{k\neq j}\|a_j - a_k\|).$$

Or, in other words, the subset of words for which the *most* affectively similar pair of words between them is as *dissimilar* as possible. Given that finding $\hat{A}$ according to this definition is intractable for large values of K, we propose to find an approximation using Algorithm 1.

Algorithm 1 Word selection

Initialization

Let A_0 be a random subset of A, of size K
$\hat{A} \leftarrow A_0$

Search

for $b \notin A_0$
  $\hat{a} = \mathrm{argmin}_{a\in\hat{A}}\|a - b\|$
  $m_{\hat{a}} = \min_{a\in\hat{A}\setminus\{\hat{a}\}}\|a - \hat{a}\|$
  $m_b = \min_{a\in\hat{A}\setminus\{\hat{a}\}}\|a - b\|$
  if $m_b > m_{\hat{a}}$
    $\hat{A} \leftarrow \hat{A} \cup \{b\}\setminus\{\hat{a}\}$
  end if

We can then use the words associated to $\hat{A}$ to present the user with options for modifying the prompt, either through a traditional UI or through a language model.

Image-driven modifiers

We consider the problem of a user who wants their generated image to be more like another target image they have seen, but in a very specific way that may not be obvious to them or easy to put into words.

Let $x_0 \in [0,1]^{3\times M\times M}$ be the (pixel) matrix associated to the current state of the generated image, and let $x_t \in [0,1]^{3\times M\times M}$ be the target image. Then, the problem can be stated as sampling from a distribution

$$\pi(x|x_0, h(x_t)),$$

where h is a feature extraction function that should isolate the aspects of the image on which the user is actually trying to condition the output. In order to discern the aspect the user is seeking to imbue in x, we can use the CLIP image encoder f (as well as the corresponding text encoder g) to figure out which words can be associated with x_t but not with x_0. That is, given a large set of available words Y, we seek a subset maximising

$$\langle g(y), f(x_t)\rangle - \langle g(y), f(x_0)\rangle, \quad (2)$$

w.r.t. y. By presenting the user with a set of words $\hat{Y} \subset Y$ maximizing Equation 2, we can get their choice, and then use it as conditioning input, thus making human decision a component of h.

From here on, we can use $\hat{y} = h(x_t)$ as the conditioning input. Alternatively, if the generative model has joint latent space for images and text, build the conditioning input as the projection of the latents as follows

$$z \doteq f(x_t) \cdot g(\hat{y}) \frac{g(\hat{y})}{\|g(\hat{y})\|}.$$

It is timely to mention that we can combine this proposal with Algorithm 1 by taking A as the set of affect scores associated to $\hat{Y}$. The suggestions presented to the user would thus become an affectively diverse subset of the descriptive words that matched the target image but not the current one. While we have not yet conducted any user studies with these techniques, this could help providing a more varied set of suggestions should the set of available words Y contain too many synonyms.

Results

For the following experiments we used the adjectives from the word dataset developed in (Warriner, Kuperman, and Brysbaert 2013), which contains word classifications into nouns, adjectives or verbs, and their corresponding affect scores.

Specificity enhancement

For the first experiment we tested enhancement suggestions provided by Algorithm 1 with five different prompts. The obtained results are described as follows, in the format that an interface may use to propose the suggestions.

- A puppy? Do you mean pure, dorky or pudgy?
- A meal? Do you mean healthy, appetizing or nutritious?
- A chair? Do you mean random, disabled or quick?
- A dragon? Do you mean righty, gorgeous or beastly?
- A king? Do you mean sensible, solid or powerful?

Some of these suggestions may be considered very good to help narrowing down the user's intentions regarding the output, while some others may be a little strange. Nonetheless, surprise is a good indicator of the creative potential of an interaction, and may lead to explore new possibilities.

In order to illustrate the complete process, we took one of the prompts and suggestions and used Stable Diffusion (Rombach et al. 2022) to generate some samples, shown in Figure 1. It can be seen that adding each suggestion does steer the drawing in a distinctive direction while retaining at least some aspects of the original image. The degree to which this is useful awaits further evaluation, but to the authors the dorky puppies are at least a bit dorky and the pudgy puppies are at least a bit pudgy.

Image-driven modifiers

In order to test the proposal of modifications through suggestions derived from a target image, we picked image pairs of objects in the same category, and produced five suggestions using Equation 2. Some of the results are shown in Figure 2, where there are two observations to be made.

Firstly, the suggestions seem pertinent and do reflect characteristics of the target images not observed on the current one.

Secondly, we compared the results obtained after modifying the prompt using one of the words suggested by our method (under "Image-driven prompt modifier") and those of guiding Stable Diffusion with two images (under "image mixture"). The naïve approach of setting both images as targets has, in the dog example on the left, introduced some unwanted changes (such as the background flowers) along with visual changes such as lightening the dog's fur. In the chair example the direct image mixture appears to have performed better, perhaps due to the absence of background detail. On the other hand, guiding the generation process by introducing a specific characteristic to the prompt, derived from the target image, results in changes much more aligned with that aspect. In the dog example on the left it is clear that "fluffiness" has increased without (significantly) altering the dog, pose, or setting. The chair results are somewhat more mixed, although here perhaps the task was harder, as antique chairs are typically not visually similar to modern moulded plastic ones.

Conclusions

In this work, we focus on the ability of a co-creative system based on generative AI to make diverse suggestions and aid its user in the task of iterative prompt exploration. We have proposed two different methods to do so in the prompt space, one based on affective modelling of words, and one based on extracting target aspects from images. Additionally, early experimental results were presented, highlighting the potential value of co-creative prompt suggestions.

It is worth mentioning that the proposed approaches for prompt improvement are generator-agnostic, meaning that they can be used with any prompt-based generative model. This constitutes a considerable asset given the rate at which new generative models are being developed.

Finally, there is still much work to be done regarding user testing. On one hand, interaction design work will be required to determine effective ways of presenting the options to users. On the other hand, we have only begun exploring the reach and limitations of these approaches.

Acknowledgments

We would like to acknowledge the Australian Research Council for funding this research (ID #DP200101059).

References

Ibarrola, F.; Lulham, R.; and Grace, K. 2023. Affect-conditioned image generation. *arXiv preprint arXiv:2302.09742*.

Liu, V., and Chilton, L. B. 2021. Design guidelines for prompt engineering text-to-image generative models. *arXiv preprint arXiv:2109.06977*.

Osgood, C. E.; May, W. H.; Miron, M. S.; and Miron, M. S. 1975. *Cross-cultural universals of affective meaning*, volume 1. University of Illinois Press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from

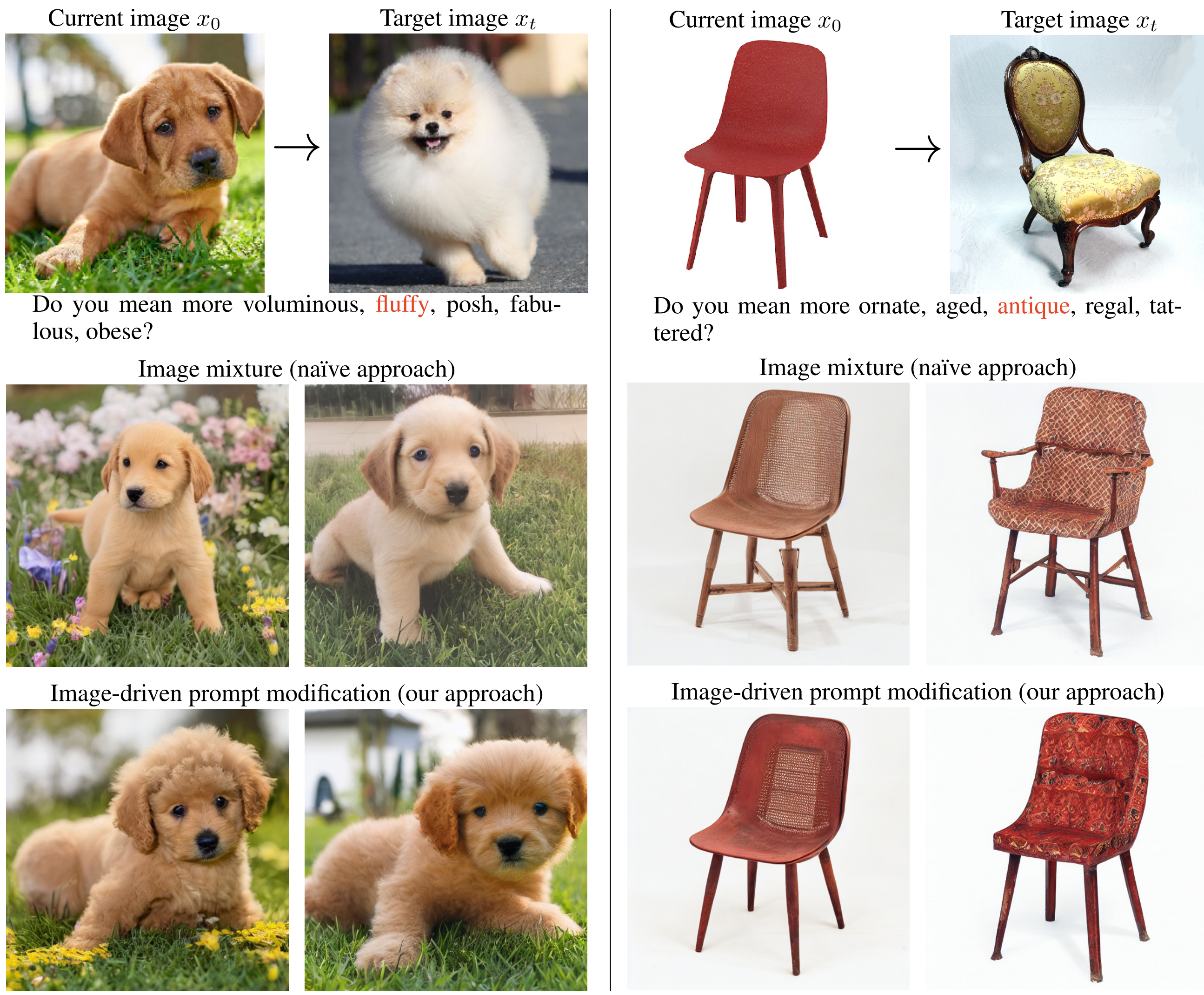


Figure 2: Two examples of the image-driven prompt modification process with Stable Diffusion. The top row shows the current image and a target images, along with the modifiers suggested by the system when presented with this image pairs. Under "Image Mixture", we show the results obtained with the original prompt and the two images as simultaneous targets. Under "Image-driven prompt modification", the results obtained with the current image and the prompt modified according to the highlighted suggestion.

natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35:36479–36494.

Warriner, A. B.; Kuperman, V.; and Brysbaert, M. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods* 45:1191–1207.

CardLab: A Simple Co-Creative Interface for Designing and Testing Cards in Hearthstone

Alexander Elton-Pym and **Kazjon Grace**
Designing with AI Lab
The University of Sydney
[alex.elton-pym, kazjon.grace]@sydney.edu.au

Abstract

Designing content for multiplayer competitive strategy games, such as collectible card games, is a complex process. Creating original, fun, and balanced content can be particularly challenging in real-world game contexts. This paper presents a pilot study of CardLab, a user-friendly creative interface for card creation and testing in (a constrained version of) the digital card game *Hearthstone*. Our study explores how designers responded to the system's feedback based on simulated games. CardLab aims to help designers more rapidly create high-quality cards, while reducing the need for extensive playtesting.

Introduction

Designing new content is a regular and essential component of many competitive strategy games, especially collectible card games (CCGs). Players expect new content to be regularly released to keep the game fresh, offer new strategic challenges, and shake up existing metagames (Carter, Gibbs, and Harrop 2012). However, keeping this new content original, fun, and balanced is a time-consuming and challenging task, often requiring designers to spend countless hours playtesting new content. As more content is released over the life of a game, generating original ideas becomes increasingly difficult. For example, in Blizzard's popular CCG *Hearthstone* (2014), over 300 new cards are released each year, requiring designers to work quickly to design, develop, and test each card before public release. Balancing content can be particularly challenging as the ecosystem of existing strategies grows, forming exponentially more interactions, any of which could lead to an unintentionally powerful combination. While balance is only one component of what makes a game fun, it is critical in competitive strategy games (Hoover et al. 2020).

As an approach to simplifying this huge and ongoing design problem (specifically in *Hearthstone*, but in concept for any competitive strategy game), we propose using a simulation engine to automatically playtest new cards, coupled with a simple user interface that allows designers to rapidly prototype them. Currently, designers and players regularly rely on statistical inference from game replay data to determine the performance of cards. This data exists only for existing cards, not new and proposed ones. Our system provides accurate simulation data accompanied by a variety of additional behavioural variables, such as the change in average game length. Our simple visual interface enables designers to prototype cards quickly without any need for coding, and the simulation engine allows for detailed feedback on their efficacy and characteristics. Card simulation is a computationally expensive task, and the feedback is only available after several hours — but this is still orders of magnitude less than it would take to playtest with humans. In one sense, our approach is intended as a speculation on the future of simulation-assisted co-creativity: to what degree can large-scale game simulation help game designers with creative tasks? We propose that this more-rapid feedback loop between design, analysis, and iterative re-design, forms a new kind of co-creative system for game content design that will become only more effective as the cost of compute decreases.

We conducted a pilot study with several *Hearthstone* players, each of whom designed a set of nine cards, which were in turn tested through 3,000 simulated games on the popular *Hearthstone* simulator *Spellsource*[1]. In a follow-up session, each user received feedback on the performance and behaviour of their cards and had the opportunity to make modifications to their designs. We analysed the participants' responses to feedback on each card's behaviour, as well as how they accordingly modified their designs.

Background

Competitive strategy games are an age-old form of entertainment, competition, and research. Games like chess have been used to benchmark human and computer capabilities since their inception. The vast majority of these games change very slowly, if at all over the years. For example, our modern version of chess has mostly remained unchanged since its 10th-century origin. This contrasts starkly with modern video games, where increasingly, game studios are following a live service model where games are continuously updated with new content throughout their lifecycle (Dubois and Weststar 2022). In this live service model, retaining customers is key to the economic success of the game, and providing a steady stream of new content while maintaining game balance is often seen as one of the most

[1] Berman and Gale, github.com/hiddenswitch/Spellsource

important factors for player's experience (Adams 2014).

Computationally creative systems have been shown to be able to assist game designers in creating balanced, fun, and original content for their games. Tanagra is a key early example of mixed-initiative creative game content design, constructing platformer levels with a focus on validating playability and comparing different generators for their expressivity (Smith, Whitehead, and Mateas 2010). Sentient Sketchbook (Liapis, Yannakakis, and Togelius 2013) is another creative interface that enables designers to collaborate with an AI to design levels for real-time strategy games. Baba is Y'all (Charity, Khalifa, and Togelius 2020) is notable in our context for incorporating automated playtesting.

Blizzard's collectible card game *Hearthstone* (2014) is – like many large, popular, continually updated games – a highly complex design domain. Evidence for this can be found in the many volatile online discussions about the relative power of new content. Research on Hearthstone has included developing adversarial agents (Świechowski, Tajmajer, and Janusz 2018) and balancing existing metagames (de Mesentier Silva et al. 2019). Like its spiritual grandparent game *Magic: The Gathering*, the space of possible *Hearthstone* decks is astronomically large, making it computationally prohibitive. The scale of this challenge has led to many simulation, analysis, and archive-exploration tools being developed by both the player community and academic researchers (Dockhorn and Mostaghim 2019). Of particular note are experiments in using neural surrogate models to predict game outcomes in reduced time (compared to extensive simulation) (Zhang et al. 2022), but it remains to be seen whether such surrogate approaches could be extended to work with new cards.

CardLab prototype

Our prototype creative interface, CardLab, enables designers to rapidly create simple *Hearthstone* cards (Figure 1). To simplify the space of required simulations (but still keep with the complexity and spirit of *Hearthstone*), we posit a "miniature" version of the game, consisting of only basic cards from the classic set. We further restrict this version to include only Hunter, Warrior, and Mage classes and require all decks to consist of 15 pairs of cards; we designate this format "classic lowlander". In CardLab, users are able to design minion cards with keywords, custom stats (mana cost, attack, and health), and simple "battlecry" effects. This simple prototype allows us to test cards in a "mini-metagame" of the three basic class decks from the game's practice mode.

After cards are converted into a format readable by *Spellsource* and inserted into the basic decks, they are simulated in 1000 games each versus Hunter, Warrior, and Mage. Simulating fewer games was observed to potentially obscure the effects of subtle card changes, while simulating more did not tend to reveal more effects. With this many games, each card takes around 40 minutes to simulate on a single 2022-era high-end workstation, making it infeasible for the interface to provide feedback online, necessitating a follow-up session. Our experiments with high-performance computing indicate that it's feasible to provide live-updating feedback after only a minor delay. We select the three most statistically significant behavioural statistics (using a T-test comparing against baseline decks) along with winrates (using Bernoulli trials) and present these to the user.

Figure 1: The CardLab interface, enabling designers to make comparisons to existing cards. A version of CardLab is hosted at `hearth-mici.web.app`

Study protocol

The user study consists of two half-hour sessions. In the first session, users design three cards for Hunter, Warrior, and Mage, while describing their design choices and thought process. Users' level of expertise with card games is determined by asking them to describe their history with *Hearthstone* and other CCGs. Throughout the study, users are prompted with questions such as "What role do you see this card filling in a deck?" to assist them in thinking aloud.

In the second session, conducted 1-7 days later, users received statistics on each card's performance in simulated games. The simulated decks are direct copies of *Hearthstone's* basic decks, with 2 copies of a vanilla neutral (i.e. non-class-specific, low-cost, no special abilities) minion selected for substitution. Cards are simulated with the *Spellsource Hearthstone* simulator, a popular java-based simulator. Games are played using a default AI from *Spellsource* which uses a form of the Minimax algorithm. This heuristic scores the hypothetical game state that would result from taking each possible move, with a policy that has been optimised with an evolutionary approach.

We asked each user to comment on their cards' performance and behaviour after receiving the feedback from the simulator, and if the results were expected or surprising. We also asked if and how they wanted to modify their cards, categorising their choices as no modification, minor modification, or significant modification. We conducted a thematic analysis of the think-aloud and post-session interviews in order to explore the design motivations of our users and the way they were affected by the simulation results. A large language model (GPT 3.5) was used (in parallel with human coding) as a supportive aid in the first pass of coding, but the final decision for all categories was human.

Figure 2: An example of the reported simulation results, included are the winrates and the most significant behavioural statistic for each match-up.

Results

We conducted our user study with 8 total users. These users have a variety of levels of experiences with *Hearthstone* and other CCGs. Two users were novices who played through the *Hearthstone* tutorial and for an hour with the basic Hunter, Warrior, and Mage decks. Four users were intermediate players who had moderate experience, with most playing when *Hearthstone* was first released. Two users had extensive experience with *Hearthstone*, having played during multiple expansions as well as experience with other CCGs like *Magic: The Gathering* and *Legends of Runeterra*. Here we present both a thematic analysis of their design motivations during the card design/re-design task, as well as an analysis of how they responded to our system's simulation results predicting their cards' performance and behaviour.

Design motivations

In this section, we describe the key design motivations that users considered important when designing cards which we identified during thematic analysis. Taken together, these motivations help us understand the creative task of card design for *Hearthstone*, and may help shape the design of future co-creative systems in that space.

Class-themed Design: Keeping the cards within the theme of the respective classes: considering class-specific abilities, and designing cards that fit within existing archetypes associated with each class.

Synergy-themed Design: Designing cards that work well together. Designers consider how cards might combine together to be more powerful than they might be individually.

Role-themed Design: Designing cards to serve a clear role in a deck. Users described their cards as being either aggressive or defensive, or designed for early- or late-game play.

Balance and Experience: The desire for cards to perform fairly and lead to a fun user experience. Users describe the kind of impact they want their cards to have on a game, adjusting power levels accordingly.

Flavour and Lore: The non-gameplay aspects of cards such as the artwork, and background lore. One user with extensive experience with *World of Warcraft*, a game from the same fictional universe as *Hearthstone*, designed many cards with their favourite characters as inspiration.

Simulation results

We categorised users' responses to the simulation results, focusing on whether each card's performance (i.e. effect on winrate) and behaviour (i.e. other effects) were expected or surprising. 8 users each designed 9 cards, for a total of 72 card designs in this analysis. In terms of winrate, 34 (47%) of the cards performed as expected, while 38 (53%) were described as "surprising". Behaviour was more predictable, with 56 (78%) cards behaving as expected, and only 16 (22%) simulation results being surprising. In other words, user expectations of how a card would act on the game were relatively accurate, but their understanding of how that card would affect the winrate was no better than random chance. This effect may partially derive from the difference between simulated and actual play, this is significant supporting evidence of the utility of our approach for card design. See Table 1 for the matrix of performance and behavioural surprise.

	Performance expected	*Performance surprising*
Behaviour expected	29	27
Behaviour surprising	5	11

Table 1: Card performance surprise and behavioural surprise.

We also categorised modifications users made after receiving feedback on their cards: minor modifications (i.e. changing mana cost, attack, and health by a few points), significant modifications (i.e. modifying or adding a new effect, or large changes to the card's mana cost, attack, and health), or no modification. Out of 72 cards, users modified 29 (40%) in a minor way, 3 (4%) were significantly modified, and the remaining 40 (56%) cards were not modified. See Table 2 for the matrix of card performance surprise and modification choice. Unsurprisingly, users more frequently modified cards whose performance was surprising. This supports the notion that CardLab can drive design iteration.

	Performance expected	*Performance surprising*
Some modification	10	22
No modification	24	16

Table 2: Card performance surprise and modification choice.

Discussion

We found that our user interface enabled users to create and test basic cards successfully. Users described a wide range of motivations behind their designs, considering how their cards would fit into decks, archetypes, and classes. Users also considered the impact cards would have on gameplay, aiming for balanced, original, and fun designs. Our AI simulation-based performance feedback helped users identify how cards could be redesigned to better achieve their intended impact on games. Given the known centrality of testing and iteration on creative design, we posit that this suggests a strong utility for this kind of "simulation-based" co-creative design system. In addition to potentially being of use in the context of *Hearthstone* card design, this suggests that the co-creative card design task may be an interesting area for future computational (co-)creativity research.

Our analysis of design intent highlighted the variety of design motivations that our participants considered when creating their cards including the potential impact of the cards on games, the player experience, and the health of the overall metagame. However, we also identified motivations which may exist in tension with the desire for originality and balance, such as the desire for cards to match existing archetypes or fit with flavourful ideas. This demonstrates the complexity of the card design task, but also potentially illustrates some directions that future co-creative systems in this space might be able to pursue.

One potential limitation of our study is that some fraction of the users' surprise at the performance of their cards may have been due to the comparatively small number of decks which we simulated. Some users designed quite complex cards that would be impactful only in niche circumstances, such as in combination with two or more other cards, or in specific deck archetypes. It is unlikely that the relatively simple player AI and card-substitution system we used in this study would showcase the strengths of such a card. Nevertheless, a significant portion of user surprise at the performance of their proposed designs appears to have arisen from a genuine expectation mismatch caused by the inherent complexity of balancing a new card in a game like *Hearthstone*. This kind of performance feedback often caused users to reconsider their card design.

Behaviour was more predictable, with many cards resulting in obvious changes to the overall behaviour of a deck (e.g. healing minions leading to more healing done). However, users described behavioural feedback as valuable, helping them better understand the impact their cards would have on games. When surprise was elicited by the behaviour (rather than performance) of proposed cards in our simulated games, it tended to initially exhibit confusion, since the changes were often indirect or secondary impacts of the proposed change. While this kind of surprise was relatively rare in our study (compared to unexpected performance), they indicate moments where the system was able to highlight complex downstream consequences the user might not otherwise have spotted. These surprises led to significant verbal reflection, as well as occasional substantial modifications. We interpret these early signs of reformulation as preliminary evidence of CardLab's capacity to facilitate co-creativity through automated playtesting.

Users sometimes described cards that they wanted to create, but could not due to limitations of our prototype. For example, many users desired more control over summoned minions, such as being able to make a card that summons a particular creature type (e.g. "Battlecry: Summon a 1/1 *Murloc*"). Other users identified a desire to have more control over the specific targeting of effects (e.g. "Destroy all *damaged* minions"), or the ability to invert a selection (e.g. "Destroy all *non*-beast minions"). The simulation engine used in our study would be able to incorporate these effects with ease, the only requirement would be for a more complex card creation user interface.

Overall, however, users found reflecting on the simulation data engaging and useful to their design process. Performance feedback allowed our users to get a better understanding of how their cards could be possibly balanced while behavioural feedback facilitated a greater understanding of card impact. While the CardLab prototype is just that – an initial exploration of the possibility of simulation-based automated playtesting – we believe it has shown the promise of this approach to co-creative game content design.

Future work to develop CardLab's capabilities could explore the system generating original cards or suggested changes to proposed designs automatically. By scaling up the simulations using high-performance computing, it would be possible to evaluate a large range of computer-designed cards, which by implementing quality diversity algorithms could be diverse, balanced, and behave as intended by designers. We also believe that future systems should explore the deck-level and meta-level considerations of card design, factoring in the complex social dynamics which drive metagame lifecycles. Recent developments in image-generating AIs and large language models have also opened up new avenues to explore the automatic creation of non-gameplay elements of cards, such as artwork, lore, and flavour-text. Future systems may be able to design all aspects of a complete card-set and this represents many exciting research directions.

References

Adams, E. 2014. *Fundamentals of Game Design*.

Carter, M.; Gibbs, M.; and Harrop, M. 2012. Metagames, paragames & orthogames: A new vocabulary. In *Proc. of the Int. Conf. on the Foundations of Digital Games*.

Charity, M.; Khalifa, A.; and Togelius, J. 2020. Baba is y'all: Collaborative mixed-initiative level design. In *IEEE Conference on Games*.

de Mesentier Silva, F.; Canaan, R.; Lee, S.; Fontaine, M.;

Togelius, J.; and Hoover, A. 2019. Evolving the hearthstone meta. In *IEEE Conference on Games*.

Dockhorn, A., and Mostaghim, S. 2019. Introducing the hearthstone-ai competition.

Dubois, L.-E., and Weststar, J. 2022. Games-as-a-service: Conflicted identities on the new front-line of video game development. *New Media & Society* 24.

Hoover, A. K.; Togelius, J.; Lee, S.; and de Mesentier Silva, F. 2020. The many ai challenges of hearthstone. *Künstliche Intelligenz* 34.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2013. Sentient sketchbook: Computer-assisted game level authoring. In *Proc. of the Int. Conf. on Foundations of Digital Games*.

Smith, G.; Whitehead, J.; and Mateas, M. 2010. Tanagra: A mixed-initiative level design tool. In *Proc. of the Int. Conf. on the Foundations of Digital Games*.

Świechowski, M.; Tajmajer, T.; and Janusz, A. 2018. Improving hearthstone ai by combining mcts and supervised learning algorithms. In *IEEE Conf. on Computational Intelligence and Games*.

Zhang, Y.; Fontaine, M.; Hoover, A.; and Nikolaidis, S. 2022. Deep surrogate assisted map-elites for automated hearthstone deckbuilding. In *Proc. of the Genetic and Evolutionary Computation Conference*.

Bits of Grass: Does GPT already know how to write like Whitman?

Piotr Sawicki[1], Marek Grześ[1], Fabricio Goes[3], Dan Brown[2], Max Peeperkorn[1], Aisha Khatun[2]

[1] School of Computing, University of Kent, Canterbury, UK
[2] Cheriton School of Computer Science, University of Waterloo, Canada
[3] Computing and Mathematical Sciences Department, University of Leicester, UK
P.Sawicki@kent.ac.uk, M.Grzes@kent.ac.uk, Fabricio.Goes@leicester.ac.uk,
Dan.Brown@uwaterloo.ca, M.Peeperkorn@kent.ac.uk, Aisha.Khatun@uwaterloo.ca

Abstract

This study examines the ability of GPT-3.5, GPT-3.5-turbo (ChatGPT) and GPT-4 models to generate poems in the style of specific authors using zero-shot and many-shot prompts (which use the maximum context length of 8192 tokens). We assess the performance of models that are not fine-tuned for generating poetry in the style of specific authors, via automated evaluation. Our findings indicate that without fine-tuning, even when provided with the maximum number of 17 poem examples (8192 tokens) in the prompt, these models do not generate poetry in the desired style.

Introduction

The recently introduced GPT-3.5 and GPT-4 models represent significant progress over the previous versions of GPT, achieving human-like performance on many tasks that were so far unattainable to Large Language Models (LLMs) (OpenAI 2023; Bubeck et al. 2023). Among creative tasks, GPT models can write poetry (Gwern Branwen 2022). In this study, however, we are concerned with generating poetry in the styles of specific authors. In our companion paper (Sawicki et al. 2023), we have examined the same challenge of generating poetry in the style of specific authors through fine-tuning GPT-3, and the results were successful. We have also found that poetry generated from GPT-3.5 (text-davinci-003) through prompt engineering only (i.e. without fine-tuning) does not follow the style of the requested author. In here, our aim is to investigate this finding further and also to check whether GPT-3.5-turbo (ChatGPT) or GPT-4 can achieve this task through prompting only. To facilitate comparison with the above-mentioned work, we attempt to generate poetry in the style of Walt Whitman without prior fine-tuning of the GPT models, and we evaluate these poems against the original works of Whitman using the automated evaluation workflows presented in our previous works (Sawicki et al. 2022; 2023).

As a main contribution of this paper, we demonstrate that generating poetry in the style of a specific author through prompting alone (whether with zero-shot or many-shot) from GPT-3.5, GPT-3.5-turbo (ChatGPT) and GPT-4 does not produce good outcome, and therefore fine-tuning is still the recommended approach.

In the next section, we describe our experimental setup, which includes three experiments to address our research question. Then, we summarize the findings of this paper and suggest the directions for future work.

Method

In this section, we describe the methodology used in this paper. First, we visually compare the difference between poems generated through the same prompt from consecutive GPT models. Then, we examine whether GPT is able to retrieve the original poems by Whitman. After that, we describe the data used for further experiments, the evaluation process, and our core results.

Three Different Models And One Prompt

While experimenting with poetry generation from consecutive versions of GPT, we have observed that the models produce poems of increasing level of complexity and length; however, the requested style is clearly not preserved. For example, Walt Whitman's poetry does not follow the 'four lines in a stanza' structure, and does not use rhyming (Bohan 1995). The majority of poems that we generated 'in the style of Walt Whitman' do follow the 'four lines in a stanza' structure and use rhyming. This, in fact, applies to most poetry generated from GPT models (including GPT-4). Only rarely will GPT deviate from this specific structure, and even then, the style does not match that of the requested author. This applies both to zero-shot prompting (where the prompt contains only the instruction to write a poem in the style of the specific author) and few-shot prompting (where in the prompt, apart from the instruction, we provide as examples a few poems by the original author). For that matter, even in a multi-step conversation with ChatGPT (GPT-3.5-turbo) and GPT-4, when the prompt highlights that the generated poems have been in 4-line stanzas with rhyme, and that the desired output should not have this structure, the model, for the most of time, still generates 4-line stanzas with rhyme.

Table 1 shows examples of the first stanzas of the poems generated from the same prompt by all three GPT models we are examining, followed by a fragment of a poem generated from the same summary using GPT-3 Curie fine-tuned for 4 epochs on the poetry of Walt Withman. The prompting procedure for the fine-tuned model is different from the prompt presented, and is fully explained in our companion paper (Sawicki et al. 2023). The results show that there is a obvious contrast between the style of writing of the fine-tuned model and the default models.

PROMPT:

Write a poem in the style of Walt Whitman.
The poem should match the following summary:

This is a poem about having courage and strength in the face of difficulties and hardships. The speaker is confident in their own determination and power, regardless of what life throws their way. They do not fear the darkness or horrors of life, knowing that they have the strength to handle what comes their way.

GPT-3.5:

Oh how I ride and stride,
In the darkest night and night's surprise.
What trouble can come my way,
Will not break me, no matter how I dread.

GPT-3.5-turbo:

Oh, the trials and tribulations of life,
How they test our will and might,
But fear not, my dear friend,
For courage and strength are within sight.

GPT-4:

I sing the song of the sailor, the wanderer of the seas,
The one who beholds the world with eyes wide and free,
In the port of Charleston, where the ships lay anchored,
I saw a sight that stirred my soul, a vision that lingered.

Fine-tuned (4 epochs) GPT-3 Curie:

OUT of the night that covers me like a hearse,
Ambient and sorrowful as the murmurous verse of those I love best,
(Not the unreplying march of Death, whose cold speech is wholly chang'd,
But the reply, the matin song of all that is, and ever shall be, in spire, in choir,
Out of the sung response, arriere glissade, in the ranks of the living,)
A vibration, blue and tremulous as a cloud of wakeful birds,
A departure and a procession stealing away into the night,
(...)

Table 1: Example of the first stanzas of the poems generated by GPT-3.5, GPT-3.5-turbo and GPT-4 from the prompt shown, followed by the fragment of a poem generated from the same summary from the fine-tuned GPT-3 Curie. The summary provided in the prompt is of the poem 'Out Of The Night That Covers Me' by William Ernest Henley. For the fine-tuned model, the prompt did not include the verbal instruction.

The style of writing is expected to change over the course of authors life. Did the young Whitman write in the same way as late Whitman? Could it be that the same author can write in a few different 'styles'? Can GPT mimic those separate 'sub-styles' with precision? Such questions are left for future research, and in here we consider the 'style' only as a very general feature, distinguishing one author's writing from another. This said, the fine-tuning workflow that we presented in (Sawicki et al. 2023) may be able to capture those more fine-grained styles, but a further analysis would be required to verify this.

Does GPT Know Whitman's Poems?

Before proceeding to poetry generation and evaluation, we first wanted to examine whether GPT is acquainted with Whitman's poetry. For that, we have run a simple experiment to check the GPT's ability to provide the complete text of requested poems.

In a sense, we are attempting to use the GPT model as a search engine here, and we are aware that, while LLMs are increasingly being used as search engines, they are notoriously unreliable at this task. Their search results are often incorrect and require verification using reliable sources (Liu, Zhang, and Liang 2023). In here, we want to accentuate the distinction between the ability to cite the text of the poems and the ability to create new poems in a requested style. The retrieved poems are compared against the ground truth, and the accuracy of the retrieved content is quantified in Table 2. These quantification can in fact support the result of (Liu, Zhang, and Liang 2023) that current GPT may return factually incorrect outputs.

This experiment is motivating the subsequent one, and our way of reasoning is as follows. The fact that a person is able to recite certain poems from memory does not imply that they are able to write in the style of that poet. For that, an average person would have to study literature, attend workshops, practice writing, etc. Our other paper (Sawicki et al. 2023) shows that fine-tuning GPT models on the works of a specific poet leads to successful acquisition of the style, similar to human studying. However, since the current GPT-4 models can generate realistic text documents in various styles that were included in its training data, a natural research question is to ask if GPT without fine-tuning has mastered the style of poets whose poems it has seen in its training data. The experiments on generating poetry without fine-tuning in the next section can in fact be seen as measuring the 'no studying' approach to style acquisition, and the fine-tuning workflow (Sawicki et al. 2023) is the 'studying' approach. In other words, if GPT-4 knows the poems of the poet in question (i.e. it has seen them in its training data and it can retrieve them when prompted), then we could expect that our experiment of generating poetry without fine-tuning would succeed in preserving the style. However, later in this paper we will show that this is not the case.

For this experiment, we have randomly selected 10 poems by Walt Whitman, and asked each of the tested GPT models to retrieve the text of the poems using the following prompt:

```
Give me the text of a poem
{TITLE OF THE POEM} by Walt Whitman.
```

Unlike in the previous versions of GPT, in GPT-3.5-turbo and GPT-4, setting the temperature parameter to 0 does not

Retrieving complete text of Whitman's poems			
Poem title	GPT-3.5	GPT-3.5-turbo	GPT-4
Spirit Whose Work Is Done	24.60%	96.05%	20.68%
Aboard At A Ship's Helm	26.43%	91.96%	94.79%
Who Learns My Lesson Complete?	21.21%	16.09%	49.59%
The World Below the Brine	28.06%	98.53%	98.53%
As At Thy Portals Also Death	27.16%	99.47%	99.47%
Eidólons	15.19%	13.82%	94.42%
I was Looking a Long While	27.60%	98.02%	98.14%
Italian Music in Dakota	24.34%	**0.0%**	82.28%
Miracles	22.81%	45.31%	67.18%
By Broad Potomac's Shore	25.05%	24.34%	23.66%
Avg. Result	**24.25%**	**58.36%**	**72.87%**

Table 2: Results of retrieving the complete text of the poems by our chosen author. The average Levenshtein distance, calculated over five trials, is utilized to quantify the similarity between the retrieved text and the original poems.

guarantee repeatability. For this reason, the process was repeated 5 times for every poem and the results were averaged. The averaged results are shown in Table 2. The similarity score reported is Levenshtein distance (Levenshtein 1966) between the original poem and the poem retrieved by the model. The Levenshtein distance is an efficient and versatile method for measuring string similarity, as it determines the minimal number of single-character edits needed to convert one string into another.

The results above 90% indicate correctly retrieved poems, with some minor differences in layout. This is acceptable, since these kind of differences are found even between different websites presenting the same poem. The lower results on GPT-3.5-turbo and GPT-4 almost always indicate that the models started to retrieve the poem correctly, but than deviated from the original text. However, the GPT-3.5 model has never correctly retrieved even a fragment of a requested poem, although these results could be different for retrieving poems by other authors. We can speculate that in the case of this model the results are always around 20% because of similar vocabulary used. It is interesting to note that in the case of "Italian Music in Dakota", GPT-3.5-turbo in all five attempts have responded: *'I'm sorry, but Walt Whitman did not write a poem titled "Italian Music in Dakota. It is possible that you are thinking of a different poet or a different poem title.'*. Therefore, we have entered 0.0% for this poem.

We can speculate that GPT's ability to retrieve the text of the poems is influenced by the number of times the poem appeared in the training dataset. Regardless, GPT-3.5-turbo and GPT-4 are, in many cases, able to retrieve the requested poems, and therefore, we can assume that those models are acquainted with the style of this poet, but as we will show later in this paper, this does not mean that they can write in the style of the requested poet, and for that—at least with the current versions of GPT models—the fine-tuning process is necessary.

Model	Version
GPT-3.5	text-davinci-003
ChatGPT	gpt-3.5-turbo (v. 2023.04.08)
GPT-4	gpt-4 (v. 2023.04.08)

Table 3: GPT versions used for poetry generation.

Experimental Setup

The principal focus of this paper is on evaluating the poetry generated through zero-shot prompts. In Reynolds and McDonell (2021) and in Kojima *et al.* (2022), it is argued that few-shot prompting is in many cases unnecessary. For example, in translation: it is not reasonable to assume that the language models can learn to translate from language A to language B just from the few examples provided in the few-shot prompt. Those works argue that the LLM already possesses the skill of (for example) translating between the two given languages, and the only purpose of the prompt is to 'invoke' that particular skill. We can speculate that this argument could extend to poetry generation using LLMs.

We were, however, intrigued by the possibility of using 8192 token-long prompts in the current version of GPT-4, which was launched 7 weeks before the submission deadline for this paper. Therefore, we also include a preliminary evaluation of poems generated from maximum-length many-shot prompts.

Data Preparation

The original author we have chosen for this work is Walt Whitman (American, 1819–1892). We use the dataset of his works created for our companion paper (Sawicki et al. 2023), which is available on our GitHub repository[1], which contains 300 poems for seven different authors (including Whitman). Since we are examining all three of the top GPT models: GPT-3.5, GPT-3.5-turbo and GPT-4 (Table 3) with zero-shot prompting, and additionally we are examining GPT-4 with many-shot prompting, we have prepared four datasets to be used in this experiment. To match the 300 samples of the original author's works, we generate 300 samples from each of the GPT models examined. For the zero-shot poetry generation, we use the following prompt for all three models (GPT-3.5, GPT-3.5-turbo and GPT-4):

```
Write a poem in the style of Walt Whitman.
The poem should match the following summary:
{SUMMARY OF THE POEM}
```

We experimented with different ways of structuring the zero-shot prompts, but have found no meaningful differences in output quality between them.

In the case of many-shot prompting of GPT-4, we generated 300 samples with the maximum possible prompt length (8192 tokens), where, apart from the instruction to generate the poem, we provided as examples 17 poems by

[1]https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation

Whitman accompanied by their summaries. The poems included in the 17-shot (i.e. 17-poem) prompt are the following: '1861', 'A Woman Waits For Me', 'Spain 1873-'74', 'Sparkles From The Wheel', 'Spirit Whose Work Is Done', 'States!', 'Tears', 'That Music Always Round Me', 'The Artilleryman's Vision', 'The Base Of All Metaphysics', 'The City Dead-House', 'The Indications', 'Aboard At A Ship's Helm', 'The Ox tamer', 'The World Below The Brine', 'These, I, Singing In Spring', and 'Think Of The Soul'. In this case the structure of the prompt is different than the one used above, to accommodate for poem examples included in the prompt:

```
These are the examples of prompts and
completions. Prompt contains the summary
of the poem, completions contains the poem
based on this summary. Write the last
completion from the prompt preceeding it,
following the examples given.
PROMPT:
{SUMMARY OF POEM 1}
COMPLETION:
{BODY OF POEM 1}
......
PROMPT:
{SUMMARY OF POEM 17}
COMPLETION:
{BODY OF POEM 17}
PROMPT:
{SUMMARY OF THE POEM TO BE GENERATED,
FROM HENLEY AND ROSETTI DATASET}
COMPLETION:
```

As before, we experimented with various ways of structuring this prompt, but found no significant differences in the output quality. One of the approaches we tried was to provide the 17-poem prompt shown above, but without the verbal instruction preceding it, thus attempting to simulate the fine-tuning process, but that did not improve the output quality.

The summaries we use for our poem generation (both zero-shot and many-shot) are taken from the dataset published in our companion paper (Sawicki et al. 2023), and these are the same summaries that were used by us for poetry generation from the fine-tuned models. These summaries were generated for poems by William Ernest Henley (1849–1903) and Christina Rossetti (1830–1894). There are 150 summaries for each author, giving 300 summaries in total. Overall, we obtain four datasets, each containing 300 poems generated from a specific GPT model as label 0, and 300 poems by the original author as label 1. Each dataset is split into training/validation subsets, with 200/100 samples per label, respectively. This two-label setup is necessary for evaluation with binary classifiers described in the Evaluation section.

When examining the dataset generated from the 17-poem prompts, we have observed that only about 25% of generated poems have deviated from the structured/rhymed style and on the surface have resembled Whitman's poetry. We can speculate that the model produces 'higher quality' outputs when prompted with a summary which is related to the subject that Whitman was writing about, and fails when we request a poem on the subject that is not present in Whitman's works, but that would require detailed analysis by the expert in English literature.

We have to stress that few-shot and many-shot prompting of GPT-4 requires a dedicated study, and in here it was treated only as a preliminary experiment.

Evaluation

Having prepared the datasets, we are fine-tuning GPT-3 models for binary classification, following the automated evaluation methodology presented in our companion paper (Sawicki et al. 2023), where evaluation is done in the following way: binary classifiers are trained on two labels, label 0 being the GPT output, and label 1 the works of the original author. If the classifier cannot distinguish between those two classes, it means that the generated poems have preserved the style/quality of the original author. On the contrary, if the classifier can distinguish between the two classes, it means that generated poems do not match the style/quality of the original author. Achieving a 50% score would mean that both labels are indistinguishable to our classifiers, which is the desired outcome.

This approach, however, comes with a caveat because it can be argued that when the evaluation results are approaching 50%, instead of indicating the successful replication of the desired style, it may simply mean that the classifier is of poor quality. For that reason, in our other paper (Sawicki et al. 2023), we have conducted a number of experiments to establish the accuracy of fine-tuned GPT-3 models as classifiers. We found them to achieve a nearly 100% accuracy, regardless of whether the two classes represented very dissimilar texts, like Whitman's poetry vs. fragments from the book on machine learning, or more similar texts, like Whitman's poetry vs. Rudyard Kipling's poetry. In there, we have also found that of the four GPT-3 models that are available for fine-tuning (the default versions of: Ada, Babbage, Curie and Davinci), the highest performing one was GPT-3 Babbage, and therefore this model was chosen as a basis for fine-tuning the classifiers in this work.

The results of classification on all four generated datasets are shown in Table 4. The table additionally includes the results from the best performing fine-tuned GPT-3 model for Whitman's poetry (FT-GPT-3 Curie 4 epochs) from (Sawicki et al. 2023). We can compare our fine-tuned models' results with the current results because of the matching setup, i.e., we used the same dataset of Whitman's works, our evaluation setup contained the same amount of samples per label, the training/evaluation split was the same (200/100), and the poems were generated from the same set of summaries.

The results show that the classifiers were able to distinguish the GPT-generated poems from the original authors' works with almost 100% accuracy. This shows that the poems generated through prompting only do not match the style/quality of writing of the original authors, while the poems generated from the fine-tuned GPT-3 models (Sawicki et al. 2023) are approaching the style/quality of the original authors' works.

These results should be interpreted with caution in the light of the fact that the binary classifiers used are entirely black-box systems, i.e. we do not know how the classification was performed. Further research is needed to address this problem and to decipher the features that lead to high

GPT-x vs Walt Whitman original			
Model	Correct	Incorrect	Accuracy
GPT-3.5	200	0	100%
GPT-3.5-turbo	200	0	100%
GPT-4	200	0	100%
GPT-4 17-poem prompt	199	1	99.5%
FT-GPT-3 Curie 4e	**123**	**77**	**61.5%**

Table 4: Results of our experiments where GPT-generated poetry is compared against the Walt Whitman's original works. Entries in the first column indicate which GPT model's output was evaluated against the Whitman's works.

similarity of the poems that have the same style according to the classifier. However, knowing that fine-tuned GPT-3 models are reliable as binary classifiers, we can, to some extent, rely on these results. Further investigation, especially including human evaluations, is necessary to thoroughly determine the quality of the GPT-generated poetry.

Future work

In the future work, we plan to analyze GPT's ability to write poetry in the 'style' of other poets, especially those who use a structured and rhymed way of writing, as this is closer to the default style of GPT-generated poetry and may yield better results.

We are also intrigued by the question of which styles of writing can be reproduced from prompt engineering alone, and at which point the fine-tuning process becomes necessary.

Now, that the models with very large context window become available (8192 tokens for current version of GPT-4, and 32K for the upcoming version), we should investigate in detail to what extent the 'few-shot' prompt engineering can improve the models' ability to generate poetry in a requested style

Conclusion

In this study, we have examined the poetry generation ability of GPT-3.5, GPT-3.5-turbo and GPT-4 when used with prompting only. We have found that the generated poems do not match the style/quality of the works of the original author, whereas the fine-tuned model can consistently reproduce the complex style of an author like Whitman. It remains to be seen whether later versions of GPT will render the fine-tuning process obsolete (for the purpose of generating poetry in the style of a specific author), but as of now, using prompting of default GPT models does not produce good results, and fine-tuning is a recommended approach.

Acknowledgments

We thank the anonymous reviewers for their feedback and suggested references.

The work of DB is supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada. MP is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds.

Author contributions

Experimental design: PS with MG, FG, DB, MP; experimental implementation: PS; writing: PS with MG, DB, FG, editing: MG, DB, FG, MP, AK.

References

Bohan, L. R. 1995. Whitman and the poetic form. In Greenspan, E., ed., *The Cambridge Companion to Walt Whitman*. Cambridge University Press. 166–193.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Gwern Branwen. 2022. GPT-3 creative fiction. https://gwern.net/gpt-3.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, 707–710. Soviet Union.

Liu, N. F.; Zhang, T.; and Liang, P. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

OpenAI. 2023. GPT-4 technical report.

Reynolds, L., and McDonell, K. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

Sawicki, P.; Grzes, M.; Jordanous, A.; Brown, D.; and Peeperkorn, M. 2022. Training GPT-2 to represent two romantic-era authors: challenges, evaluations and pitfalls. In *Proc. of ICCC*. Association for Computational Creativity (ACC).

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; Khatun, A.; and Paraskevopoulou, S. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *Proc. of ICCC*. Association for Computational Creativity (ACC). https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation/.

Steps Toward Quantum Computational Creativity

Liane Gabora
Department of Psychology
University of British Columbia
liane.gabora@ubc.ca

Abstract

At the heart of creativity is the forging of new concept combinations and the adapting of existing ideas to new situations. However, these processes have resisted mathematical description; concepts violate the rules of classical logic when they interact, e.g., concept combinations can exhibit emergent features not possessed by their constituent concepts. These challenges can be addressed using the quantum cognition framework, wherein nonclassical behavior is described in terms of superposition, entanglement, and interference. While in classical probability theory events are drawn from a common sample space, in quantum models events are defined only with respect to a measurement, or (in quantum cognition) a context, and the probabilities reflect the underlying reality. The measurement (or context) causes collapse from a superposition state to a definite eigenstate. The paper explains how creativity can be modeled using quantum cognition approach with an illustrative example, and discuss how the approach could be implemented computationally. Quantum computing is widely expected to revolutionize many fields in the near future through immense increases in speed and computing power. The time may be ripe to explore the potential of quantum computational creativity.

Introduction

Though creativity is a vast and multifaceted topic (Jordanous and Keller 2016), at its core is the generation of new concept combinations (Estes and Ward 2002). However, modelling concept combination turns out not to be straightforward; there is extensive evidence that people use conjunctions and disjunctions of concepts in ways that violate the rules of classical (including fuzzy) logic; *i.e.*, concepts interact in ways that are non-compositional (Aerts, Aerts, and Gabora 2009; Estes and Ward 2002; Hampton 1988; Osherson and Smith 1981). This noncompositionality is observed in *exemplar typicalities* (*e.g.*, although people do not rate 'guppy' as a typical `PET`, nor a typical `FISH`, they rate it as a highly typical `PET FISH`), as well as *properties* (*e.g.*, although people do not rate 'talks' as a characteristic property of `PET` or `BIRD`, they rate it as characteristic property of `PET BIRD`). This problem has made concepts resistant to mathematical description, and plagued efforts to model how new meanings emerge when people combine concepts and words into larger semantic units such as conjunctions, phrases, or sentences.

One study of this phenomenon analyzed data on the relative frequency of membership of specific exemplars of general categories or concepts, as well as of conjunctions of them (Hampton 1988). For example, participants were asked whether an exemplar such as `Mint` is a member of `FOOD`, whether it is a member of `PLANT`, and whether it is a member of `FOOD AND PLANT`. For several items, participants assessed the examplar as more strongly a member of `FOOD AND PLANT` than of either of the two component concepts `FOOD` and `PLANT` alone. The relative frequency of membership for `Mint`, for example, was 0.87 for the concept `FOOD`, 0.81 for the concept `PLANT`, and 0.9 for the conjunction `FOOD AND PLANT`. It is difficult to conceive of a classical probability model that could encompass this finding, and it was proven that no such model exists, but that it is possible to describe this using a quantum probability model (Aerts 2009). Findings such as this suggest that a quantum approach could prove useful in computational creativity.

This paper outlines the rationale for a quantum approach to modeling creativity, and illustrates the approach using a specific example. It then discusses possible ways to computationally implement the approach. A glossary of terms is provided at the end. Note that the quantum approach does not assume that anything at the quantum level of subatomic particles affects cognition (and in this sense, it is somewhat unfortunate that it has come to be called the quantum approach). It merely uses a generalization of mathematics that was first applied to quantum mechanics.

Rationale for the Quantum Approach

Mental states involving uncertainty, ambiguity, and contextuality figure prominently in creative cognition, and quantum formalisms are uniquely suited to the formal description of such states. This is because a quantum system can be in a *superposition state*, which has the potential to transition into, or (in the quantum jargon) *collapse to* different states depending on how it is measured, or (in quantum cognition), the perspective, or *context*, from which it is considered. For example, consider the situation in which a farmer wonders what to do with an old tire. If he encounters a horse, he might consider the concept `TIRE` in the context <u>see horse</u>, which might lead him to invent a `TIRE BOWL`, i.e., a bowl for his horse (Figure 1). However, if he considers TIRE in the context see child, he might be more likely to make a TIRE SWING. Much as a qubit is not in a specific state (neither 0 nor 1) until a gate causes it to collapse to either 0 or 1, the mental state of wondering what to do with the old tire encompasses multiple possible ideas for reuses of the old tire, and the context influences its 'collapse' to one of them or another. Since the superposition state of a concept incorporates these different possible contexts and outcomes, the quantum approach appears to be better-suited than classical approaches to capture the open-endedness of creativity.

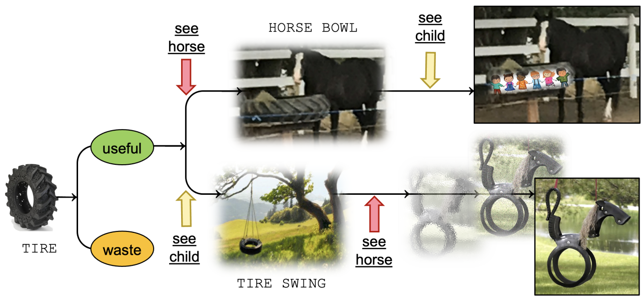


Figure 1: When the context see horse comes before see horse, the old tire is more likely to be used as a horse bowl (top). When the order of contexts is reversed, it is more likely to be used as a horse-shaped tire swing (bottom).

Accordingly, formalisms first used to model situations of ambiguity and contextuality in quantum mechanics (Khrennikov 2010; Wang et al. 2013). have been used to modeled many phenomena relevant to creativity, including semantic spaces and the combination of words and concepts (Aerts 2009; Aerts and Gabora 2005; Gabora and Aerts 2002; Bruza et al. 2009; 2015; Clark, Coecke, and Sadrzadeh 2008; Coecke, Sadrzadeh, and Clark 2010; Lewis, Marsden, and Sadrzadeh 2020), similarity and memory (Pothos and Busemeyer 2022; Nelson et al. 2013), information retrieval (Van Rijsbergen 2004; Melucci 2008), decision making and probability judgement errors, including order effects (Busemeyer, Wang, and Townsend 2006; Busemeyer et al. 2011; Mogiliansky, Zamir, and Zwirn 2009; Yukalov and Sornette 2009) language and text perception (Aerts and Beltran 2020; Surov et al. 2021), cultural evolution (Gabora and Aerts 2009; Gabora, Scott, and Kauffman 2013), tonal attraction in music (Beim Graben and Blutner 2019), and humor (Gabora and Kitto 2017). There have also been findings that cognitive processes exhibit signature features of quantum structure such as superposition, entanglement, and interference (Aerts 2009; Aerts et al. 2012; 2016; Busemeyer and Bruza 2012; Surov et al. 2019).[1]

Brief Outline of the Quantum Approach

Before applying the quantum approach to creativity, we briefly outline how quantum probability differs from classical probability. Classical probability describes events by considering subsets of a common sample space (Isham 1995). That is, considering a set of elementary events, we find that some event e occurred with probability p_e. Classical probability arises due to a lack of knowledge on the part of the modeller. The act of measurement merely reveals an existing state of affairs; it does not interfere with the results. In contrast, quantum models use variables and spaces that are defined (sometimes implicitly) with respect to a particular measurement. Measurements (or contexts) directly influence quantum systems, imposing definite states that may not have been present beforehand (Freedman and Clauser 1972).

In the quantum formalism, the *state* Ψ representing some aspect of interest in our system is written as a linear superposition of a set of possible states referred to as *basis states* $\{\phi_i\}$ in a *Hilbert space*, denoted $\mathcal{H}$, which allows us to define notions such as distance and inner product. In creating this superposition, we weight each basis state with an amplitude term, denoted a_i, which is a complex number representing the contribution of a component basis state ϕ_i to the state Ψ. Hence $\Psi = \sum_i a_i \phi_i$. The probability that the state changes to that basis state upon measurement is $|a|^2$. This non-unitary change of state is called *collapse*, which is modeled as a projection.

The choice of basis states is determined by the value being measured, termed the *observable*, $\hat{O}$. In quantum mechanics, the observables are physical quantities such as position or momentum values (but as we shall see, in quantum cognition they can be, for example, specific instantiations of a concept in a particular context). The potential measurement outcomes o_i correspond to states of the entity of interest. These resultant states of our measurement (or context), are the basis states of the Hilbert space, thus they shape how we model the entity to be measured, and its possible outcomes o_i. The basis states corresponding to an observable outcome are referred to as *eigenstates*. Observables are represented by operators.[2] Upon *measurement*, the state of the entity is projected onto one of the basis states.

It is also possible to describe combinations of two entities within this framework, and to learn about how they might influence one another, or not. Consider two entities A and B with Hilbert spaces $\mathcal{H}_\mathcal{A}$ and $\mathcal{H}_\mathcal{B}$. We may define a basis $|i\rangle_A$ for $\mathcal{H}_\mathcal{A}$ and a basis $|j\rangle_B$ for $\mathcal{H}_\mathcal{B}$, and denote the amplitudes associated with the first as a_i^A and the amplitudes associated with the second as a_j^B. The Hilbert space in which a composite of these entities exists is given by the tensor product $\mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$. The most general state in $\mathcal{H}_\mathcal{A} \otimes \mathcal{H}_\mathcal{B}$ has the form

$$|\Psi\rangle_{AB} = \sum\nolimits_{i,j} a_{ij} |i\rangle_A \otimes |j\rangle_B \tag{1}$$

This state is separable if $a_{ij} = a_i^A a_j^B$. It is inseparable, and therefore an entangled state, if $a_{ij} \neq a_i^A a_j^B$.[3]

[1]The quantum approach may be related to the signal processing approach to meaning generation, and hence creativity, based on spectral modelling of brain activity (Wiggins 2020).

[2]Specifically, Hermetian operators, which are defined on a complex inner product space, but we do not go into that here.

[3]It has been argued that the quantum field theory procedure, which uses Fock space, gives a superior internal structure for modelling concept combination (Aerts 2009). Fock space is the direct sum of tensor products of Hilbert spaces, so it is also a Hilbert space. For simplicity, we omit such refinements here.

Quantum Cognition and its Application to Creativity

We first outline in general terms how the quantum framework is adapted to cognition, and then apply it to creativity using a specific example. The set of possible states of a mental construct, such as a particular concept, is given by Σ. The amplitude term associated with a basis state represented by a complex number coefficient a_i gives a measure of how likely a given change of state is. The basis states represent possible instantiations of the concept. States are represented by unit vectors, and all vectors of a decomposition have unit length, are mutually orthogonal, and generate the whole vector space, thus $\sum_i |a_i|^2 = 1$. Self-adjoint operators[4] are used to define context-specific subspaces. The context causes the state of concept to collapse to one of its eigenstates. The role of the observable is played by the detectable changes to the . Thus, we model change in the concept under a specific context by collapse to a new state.

Each possible form of a concept represented by a particular basis state can be broken down into a set $f_i \in \mathcal{F}$ of features (or properties), which may be weighted according to their relevance with respect to the current context. The *weight* (or renormalized applicability) of a certain property given a specific state of the concept $|p\rangle$ and a specific context $c_i \in \mathcal{C}$ is given by ν. For example, $\nu(p, f_1)$ is the weight of feature f_i for state p. Thus ν is a function from the set $\Sigma \times \mathcal{F}$ to the interval $[0, 1]$. We write:

$$\begin{aligned} \nu : \Sigma \times \mathcal{F} &\rightarrow [0,1] \\ (p, f_i) &\mapsto \nu(p, f_i) \end{aligned} \tag{2}$$

A function μ describes the transition probability from one state to another under the influence of a particular context. For example, μ*(q, e, p)* is the probability that state p under the influence of context e changes to state q. Mathematically, μ is a function from the set $\Sigma \times \mathcal{C} \times \Sigma$ to the interval $[0, 1]$, where $\mu(q, e, p)$ is the probability that state p under the influence of context e changes to state q. We write:

$$\begin{aligned} \mu : \Sigma \times \mathcal{C} \times \Sigma &\rightarrow [0,1] \\ (q, e, p) &\mapsto \mu(q, e, p) \end{aligned} \tag{3}$$

Thus our quantum model consists of the 3-tuple $(\Sigma, \mathcal{C}, \mathcal{F})$, and the functions ν and μ.

Let us now make this more concrete using the example of a farmer wondering what to do with an old tire. The state of TIRE, represented by vector $|p\rangle$ of length equal to 1, is a linear superposition of basis states in a complex Hilbert space $\mathcal{H}$ which represent possible states (instances, interpretations, or types) of this concept, including typical states such as SNOW TIRE, and atypical ones such as TIRE SWING. The different states of TIRE can be described as different subspaces into which TIRE can be projected, and thereby, experienced as meaningful. Our knowledge of the possible uses, or affordances, of TIRE comes to us by way of its projections into these subspaces.

[4]Unlike Hermetian operators, self-adjoint operators are defined over the real or complex numbers.

For simplicity, let us suppose that the farmer's initial conception of TIRE is a superposition of only two possibilities (Figure 2). The possibility that the tire is considered *useful* is denoted by the unit vector $|u\rangle$. The possibility that it should be discarded as *waste* is denoted by the unit vector $|w\rangle$. The state of the concept TIRE is denoted $|t\rangle$. Their relationship is given by the equation

$$|t\rangle = a_0|u\rangle + a_1|w\rangle, \tag{4}$$

where a_0 and a_1 are the amplitudes of $|u\rangle$ and $|w\rangle$ respectively in the farmer's mind. States are represented by unit vectors and all vectors of a decomposition such as $|u\rangle$ and $|w\rangle$ have unit length, are mutually orthogonal and generate the whole vector space; thus $|a_0|^2 + |a_1|^2 = 1$.

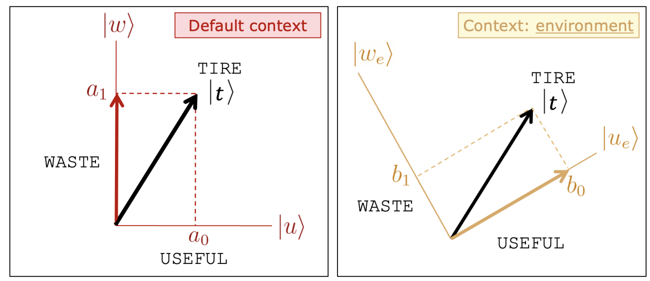


Figure 2: Left: In the default context, TIRE likely collapses to projection vector $|w\rangle$ which represents that it is waste, so $a_0 < a_1$. Right: In the context environment, it likely collapses to orthogonal projection vector $|u\rangle$ which represents that it is useful, so $b_0 > b_1$.

Note that, in someone else's mind a_0 and a_1 might be different (as epitomized in the saying "one person's trash is another person's treasure"). Indeed, if the farmer sees a recycle sign, and thinks of TIRE in the context environment, he himself may feel inspired to find a creative reuse for the tire. Consider the situation in which this is indeed what happens. If the farmer has a horse, the context see horse might be a member of the set $\mathcal{C}$ of possible contexts that could influence the subsequent state of the concept TIRE. The concept TIRE in the context see horse is denoted $|t_h\rangle$.

Activation of the set $\mathcal{L}$ of properties of TIRE, *e.g.,* the property 'weather resistant' denoted f_1, spreads to other concepts for which these properties are relevant. Possible items that must be weather-resistant that could be made from the tire, and thus possible states of $|p_h\rangle$, are (1) a bowl for the horse, or (2) a saddle. We denote HORSE BOWL and TIRE SADDLE $|l\rangle$ and $|e\rangle$, respectively, and the corresponding possible states of TIRE are denoted $|l_h\rangle$ and $|e_h\rangle$. Thus, the restructured conception of TIRE in the context of see horse is given by

$$|t_h\rangle = b_0|u_h\rangle + b_1|w_h\rangle \tag{5}$$

where

$$|u_h\rangle = b_2|t_h l_h\rangle + b_3|t_h e_h\rangle + b_4|t_h s_h\rangle, \tag{6}$$

and where $|t_h l_h\rangle$ and $|t_h s_h\rangle$ represent the possibility that he thinks of HORSE BOWL and TIRE SADDLE respectively,

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

and $|t_h\rangle$ represents the possibility that even in the context see horse the farmer thinks of the idea TIRE SWING.

Consider the set of strongly weighted properties of SADDLE, such as 'made of leather' denoted f_2 and 'has stirrups', denoted f_3. Because 'made of leather' and 'has stirrups' are not properties of TIRE, $\nu(t, f_2) << \nu(e, f_2)$, and similarly $\nu(t, f_3) << \nu(e, f_3)$. Therefore, $|b_4|$ is small. However, consider the property of bowl 'has curved edges to keep food in', denoted f_4. Since the curved edges of a tire could stop horse food from falling out, $\nu(t, f_4) \approx \nu(l, f_4)$. Therefore, $|b_3|$ is large. Thus $\mu(l, h, t) >> \mu(e, h, t)$. In the context see horse, the concept TIRE is more likely to collapse to HORSE BOWL. Note that HORSE BOWL has the emergent property, 'holds horse food,' which is a property of neither TIRE nor BOWL. We can model the emergence of new properties by describing TIRE BOWL as an entangled state of the concepts TIRE and BOWL. HORSE BOWL is thereafter a new state of both concepts TIRE and BOWL. Entanglement introduces interference of a quantum nature, and hence the amplitudes are complex numbers (Aerts 2009).

We now consider the contexts see horse and see child, and for simplicity we consider only two possible outcomes for each, HORSE BOWL and TIRE SWING. This could be depicted in an analogous manner to Figure 2, with the contexts default context and environment replaced by see horse and see child, and USEFUL and WASTE replaced by TIRE SWING and HORSE BOWL on the x and y axes respectively. Once again, the context influences the probabilities associated with each reuse idea. Interestingly, as depicted in Figure 1, if *both* contexts are encountered, the final creative outcome depends on the order in which the contexts are encountered. If see horse is encountered first, the thought trajectory likely goes the HORSE BOWL route, but if the child is encountered first, it likely goes the TIRE SWING route, culminating in HORSE TIRE SWING. Such order effects are accommodated in quantum formalism because projection to subspace a_1 then $b_1 \neq$ projection to subspace b_0 then a_0.

The TIRE example shows that a quantum cognition approach to concept interactions, which has been shown to be consistent with human data (Aerts 2009; Aerts et al. 2016), can model the restructuring of concepts during the honing of a creative idea.

Quantum Computational Creativity

Quantum cognition could be incorporated into computational creativity building on existing computational quantum cognition models. The quantum Bayesian network (QBN) combines classical Bayesian networks with quantum probability theory to represent and model human decision-making under uncertainty (Low, Yoder, and Chuang 2014). QBNs have been useful for explaining cognitive biases such as the conjunction fallacy, but more promising routes for modeling creativity are quantum machine learning (Biamonte et al. 2017) or the quantum associative memory approach (Ventura and Martinez 2000), the latter of which proposes that human memory retrieval is influenced by quantum-like interference effects that can account for context-dependent memory. It has been proposed that while such interference effects may have a disruptive effect on retrieval, they enable the fusion of seemingly unrelated context-dependent concepts and ideas that lie at the core of creativity (Gabora and Ranjan 2013). This suggests that such interference effects may be important for computational creativity.

Conclusions

This paper discussed the rationale for bridging quantum cognition and computational creativity, and outlined key steps toward the realization of such a move. A quantum computational creativity model is only as accurate as the number of basis states, properties, and contexts it includes, but with the advent of large language models, this becomes less prohibitive. The approach incorporates the ongoing interaction between potentiality (superposition state) and actualization (eigenstate), and it is this capacity of a quantum system to exist in a superposition of multiple states that lies behind the speed and power of quantum computing, and the widespread belief that it could revolutionize many aspects of our lives. It is widely believed that quantum computing will have a near-term revolutionary impact on many fields, thus, the time seems ripe for exploring its incorporation into computational creativity.

Appendix A: Definitions

Amplitude: A complex number similar to a probability value that gives the likelihood of a particular quantum state. Amplitudes can interfere (constructively or destructively).

Collapse: The change when a quantum system is measured, from a superposition of states to a single definite state.

Eigenstate: A state associated with a particular observable, or context, that has a definite value when measured.

Entanglement: A phenomenon wherein two or more quantum structures are linked—even if widely separated—such that the state of one cannot be described independently of the others, a change to one instantly affects the others.

Hilbert space: A mathematical vector space for describing quantum states and their dynamical evolution.

Interference: The phenomenon wherein waves associated with different quantum possibilities overlap and interact, either constructively, such that they amplify each other, or destructively, such that they cancel each other out.

Observable: A measurable quantity represented by a mathematical operator that acts on a quantum state.

Qubit (short for 'quantum bit'): The fundamental unit of quantum information. Unlike the classical bit, the basic unit of classical computing, which can be either a 0 or a 1, a qubit can exist in a superposition of both 0 and 1 simultaneously.

Superposition: Unlike classical systems, where objects have definite properties, a quantum system can be in a combination of different states at the same time. This combined state is referred to as a superposition.

Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant number GR026749 AWD-024382. Thanks to former student Haley Ramier for the photograph of a horse-shaped tire swing.

References

Aerts, D.; Aerts, S.; and Gabora, L. 2009. Experimental evidence for quantum structure in cognition. In *Proc. Intl. Sympos. Quant. Interact.* Berlin: Springer. 59–70.

Aerts, D., and Beltran, L. 2020. Quantum structure in cognition: Human language as a boson gas of entangled words. *Foundations of Science* 25(3):755–802.

Aerts, D., and Gabora, L. 2005. A theory of concepts and their combinations ii: A Hilbert space representation. *Kybernetes* 34(1/2):192–221.

Aerts, D.; Broekaert, J.; Gabora, L.; and Veloz, T. 2012. The guppy effect as interference. In *Proc. Intl. Sympos. Quant. Interact.*, 36–47. Berlin: Springer.

Aerts, D.; Broekaert, J.; Gabora, L.; and Sozzo, S. 2016. Generalizing prototype theory: A formal quantum framework. *Front. Psychol.* 7:418.

Aerts, D. 2009. Quantum structure in cognition. *J. Math. Psych.* 53(5):314–348.

Beim Graben, P., and Blutner, R. 2019. Quantum approaches to music cognition. *J. Math. Psych.* 91:38–50.

Biamonte, J.; Wittek, P.; Pancotti, N.; Rebentrost, P.; Wiebe, N.; and Lloyd, S. 2017. Quantum machine learning. *Nature* 549:195–202.

Bruza, P.; Kitto, K.; Nelson, D.; and McEvoy, C. 2009. Is there something quantum-like about the human mental lexicon? *J. Math. Psych.* 53:362–377.

Bruza, P.; Kitto, K.; Ramm, B.; and Sitbon, L. 2015. A probabilistic framework for analyzing the compositionality of conceptual combinations. *J. Math. Psych.* 67:26–38.

Busemeyer, J., and Bruza, P. 2012. *Quantum Models of Cognition and Decision*. Cambridge University Press.

Busemeyer, J.; Pothos, E.; Franco, R.; and Trueblood, J. 2011. A quantum theoretical explanation for probability judgement errors. *Psychol. Rev.* 118:193—-218.

Busemeyer, J.; Wang, Z.; and Townsend, J. 2006. Quantum dynamics of human decision making. *J. Math. Psych.* 50:220–241.

Clark, S.; Coecke, B.; and Sadrzadeh, M. 2008. A compositional distributional model of meaning. In *Proc. Intl. Sympos. Quant. Interact.*, 133–140. Berlin: Springer.

Coecke, B.; Sadrzadeh, M.; and Clark, S. 2010. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis* 36:345–384.

Estes, Z., and Ward, T. 2002. The emergence of novel attributes in concept modification. *Creat. Res. J.* 14:149–156.

Freedman, S., and Clauser, J. 1972. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.* 28:938–941.

Gabora, L., and Aerts, D. 2002. Contextualizing concepts using a mathematical generalization of the quantum formalism. *J. Exp. Theor. Artif. Intell.* 14(4):327–358.

Gabora, L., and Aerts, D. 2009. A model of the emergence and evolution of integrated worldviews. *J. Math. Psych.* 53:434–451.

Gabora, L., and Kitto, K. 2017. Toward a quantum theory of humor. *Front. Phys.* 4:53.

Gabora, L., and Ranjan, A. 2013. How insight emerges in distributed, content-addressable memory. In Bristol, A.; Vartanian, O.; and Kaufman, J., eds., *The neuroscience of creativity*. Cambridge, MA: MIT Press. 19–43.

Gabora, L.; Scott, E.; and Kauffman, S. 2013. A quantum model of exaptation: Incorporating potentiality into evolutionary theory. *Prog Biophys Mol Biol* 113:108–116.

Hampton, J. A. 1988. Disjunction of natural concepts. *Memory and Cognition* 16:579–591.

Isham, C. J. 1995. *Lectures on Quantum Theory*. London: Imperial College Press.

Jordanous, A., and Keller, B. 2016. Modelling creativity: Identifying key components through a corpus-based approach. *PloS One* 11:e0162959.

Khrennikov, A. 2010. *Ubiquitous Quantum Structure: From Psychology to Finance*. Berlin, Germany: Springer.

Lewis, M.; Marsden, D.; and Sadrzadeh, M. 2020. Semantic spaces at the intersection of nlp, physics, and cognitive science. *Journal of Applied Logics* 7:677–682.

Low, G.; Yoder, T.; and Chuang, I. 2014. Quantum inference on Bayesian networks. *Physical Review A* 89(6):062315.

Melucci, M. 2008. A basis for information retrieval in context. *ACM Trans. Inf. Syst.* 26:14:1–14:41.

Mogiliansky, A. L.; Zamir, S.; and Zwirn, H. 2009. Type indeterminacy: A model of the KT (Kahneman–Tversky)-man. *J. Math. Psych.* 53(5):349–361.

Nelson, D. L.; Kitto, K.; Galea, D.; McEvoy, C. L.; and Bruza, P. D. 2013. How activation, entanglement, and search in semantic memory contribute to event memory. *Memory and Cognition* 41(6):717–819.

Osherson, D. N., and Smith, E. E. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition* 9:35–58.

Pothos, E. M., and Busemeyer, J. R. 2022. Quantum cognition. *Annual Review of Psychology* 73:749–778.

Surov, I. A.; Pilkevich, S. V.; Alodjants, A. P.; and Khmelevsky, S. V. 2019. Quantum phase stability in human cognition. *Front. Psychol.* 10:929.

Surov, I. A.; Semenenko, E.; Platonov, A.; Bessmertny, I.; Galofaro, F.; Toffano, Z.; Khrennikov, A. Y.; and Alodjants, A. 2021. Quantum semantics of text perception. *Scientific Reports* 11(1):1–13.

Van Rijsbergen. 2004. *The Geometry of Information Retrieval*. Cambridge University Press.

Ventura, D., and Martinez, T. 2000. Quantum associative memory. *Inf. Sci.* 124:273–296.

Wang, Z.; Busemeyer, J. R.; Atmanspacher, H.; and Pothos, E. M. 2013. The potential of using quantum theory to build models of cognition. *Top. Cogn. Sci.* 5(4):672–688.

Wiggins, G. 2020. Creativity, information, and consciousness: the information dynamics of thinking. *Phys. Life Rev.* 34:1–39.

Yukalov, V., and Sornette, D. 2009. Processing information in quantum decision theory. *Entropy* 11(4):1073–1120.

Minimally Juxtapository Tasks as a Co-Creative Systems User Study Method

Geoffrey Lazarus and **Kazjon Grace**
Designing with AI Lab
The University of Sydney
[geoffrey.lazarus,kazjon.grace]@sydney.edu.au

Abstract

Conceptual design is an exploratory stage in the creative design process that is challenging to augment with computational techniques. Part of this challenge comes innately from the flexibility and reframing-centric nature of the task itself, but an equal contributor is the trouble of measuring, standardising, and working with conceptual design in experiments. We propose a model for studying conceptual design in co-creative systems based on minimally juxtapository tasks (MJTs). In this paper we detail a case study of conceptual designing with AI-based art tool, ArtBreeder, using our new task format. Through MJTs, participants engaged with features of ArtBreeder and reflected upon its capacity to assist them. We performed thematic analysis on post-task interviews to derive a series of themes for use in better understanding user attitudes and behaviours. The findings help frame the shortcomings of existing conceptual exploration tools, validating the MJT method.

Introduction

Conceptual design is an important stage in the design process (Pahl et al. 2007), exploring not just possible designs but constraints, requirements, and interpretations of the brief that define the emerging space of possible designs (Maher, Poon, and Boulanger 1996). Conceptual designers enact the prototypical exploration of forms, motifs, and constraints in early phases of designing. At present, there have been few computational tools used in conceptual design compared to other design activities. It appears that conceptual designing is one of the last aspects of designing to be transformed by computational tools – likely due to the requirement for representational flexibility and the regular re-invention of requirements.

Recent research on generative machine learning may offer an opportunity to develop a new form of computational technique that is more well-suited to the conceptual design stage (Maher, Poon, and Boulanger 1996). However, given the unique nature of conceptual design, no well-validated methods for exploring the efficacy of such techniques exists (Lawton et al. 2023a). In this paper we propose a generalisable task format as a step towards standardising research in co-creative systems for conceptual design. Our task format, *minimally juxtapository tasks*, or MJTs, focusses on finding the simplest possible task that incorporates traditionally opposed concepts. Tensions and juxtapositions are common antagonists in conceptual design, and their confrontation is one known driver of creativity and innovation (Dorst 2015). We define the MJT concept, then present a case study in which a generative machine learning system designed for an artistic domain is re-purposed to conceptual design tasks through their use. This study lets us explore how designers interact with tensions and juxtapositions that occur in much more complex tasks, but in the simplicity of a short user study. We believe this method offers a glimpse into how to best design future conceptual design tools.

ArtBreeder (screenshot shown in Figure 1) is a visual image synthesis tool used in the concept art domain, where artists working on games, films and other media develop early concepts for their characters, environments, and scenes. We detail a qualitative study in which 11 participants used ArtBreeder to generate character faces based on MJT-style prompts, followed by an interview exploring how the tool affected their creative process. We then conduct a thematic analysis of the results to explore the potential for minimally juxtaposed tasks for conceptual co-creative AI.

Related Literature

Conceptual Design is a preliminary stage in the design process for exploring possible solutions, requirements, interpretations and constraints in response to the design problem (Bentley and Wakefield 1997). Conceptual design involves re-representing and reformulating designs. The iterative reformulation of conceptual design can be interpreted through an exploratory co-evolution model (Maher, Poon, and Boulanger 1996; Wiltschnig, Christensen, and Ball 2013), where the problem space and solution space modify one another and refine both the design solution and design requirements.

Recent advances in generative machine learning techniques have enabled neural networks to synthesise artefacts that are indistinguishable from human-generated examples in many domains, at least under certain constraints (Bessette, Fol Leymarie, and W Smith 2019). These advances have been applied to conditional image synthesis (including sketches (Di and Patel 2017), text-to-image synthesis (Zhang et al. 2017), text-guided image editing (Li et al. 2020) and style transfer (Gatys, Ecker, and Bethge 2016).

Examples of how these techniques have been applied to creative domains can be found in DALL-E (Ramesh et al. 2021), CLIP (Radford et al. 2021), Stable Diffusion, and Sketch2GAN (Wang, Bau, and Zhu 2021). These forms of cross-modal synthesis lend to an important aspect of conceptual design: the re-representation of designs.

Research in creative systems has described a continuum of roles among humans and creative systems: creativity support tools, co-creative agents, and fully autonomous creative AI. Creativity support tools assist human creativity but do not necessarily use AI to do so. Fully autonomous systems generate creative artefacts themselves, with the involvement of humans limited to down-stream evaluation and curation of their output. Co-creativity, the focus of this paper, is instead a collaboration between humans and computers to develop shared creative artefacts (Davis 2013). Notable co-creative systems include Sentient Sketchbook (Liapis et al. 2013) in game design, the Creative Sketching Apprentice (Karimi et al. 2019) and Reframer (Lawton et al. 2023b) in sketching, EvoFashion (Lourenço et al. 2017) in clothing design, and a huge variety in the domain of music (Ford et al. 2022).

Minimally Juxtapository Tasks (MJTs)

We developed MJTs as a way to explore conceptual design in user studies through simple tasks, based on the question: how simple can a study task be while still retaining analogous enough to a real-world design problem that the appropriate cognitive machinery must be recruited to solve it? Our proposed answer to that question coalesced into the MJT, which can be defined as:

> *A creative brief, typically described in a single sentence, that requires imbuing an artefact with two concepts that are conceptually, affectively, or otherwise significantly opposed.*

The juxtaposition at the heart of these simple tasks makes for a creatively interesting brief, requiring designers/artists to negotiate ideas that would stereotypically be opposed. For example, in our face-generation case study using ArtBreeder, the four tasks were to depict:

- "A grizzled veteran with a heart of gold."
- "A sweet senior citizen with a wild side."
- "A detective with a music career side hustle."
- "A zombie politician."

Case Study Methodology

As an initial exploration of the efficacy of MJTs in a real-world co-creative systems context we used the popular ArtBreeder web-based co-creative platform as the subject of a case study. ArtBreeder[1] is a web-based platform created by Joel Simon for artistic exploration and image generation that uses machine learning techniques to create and evolve digital art. It uses a combination of generative adversarial networks (GANs) and autoencoders to manipulate images.

[1]www.artbreeder.com

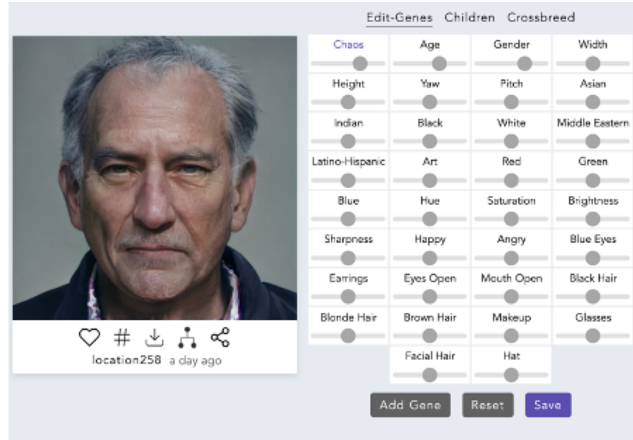


Figure 1: An example of the ArtBreeder system, showing the Edit Genes feature for a face.

We investigated the efficacy of ArtBreeder on conceptual designing using MJTs with 11 user participants. For the first three tasks, one of the system's features was used: (1) Edit-Genes, modifying distinct image attributes such as eye colour or age, through sliders; (2) Children, mutating a selected image; and (3) Crossbreed, blending two images with content and style sliders. In the final task they used all three.

The participants were practicing designers from a variety of disciplines, six women and five men. Tasks were conducted via remotely recorded video calls and screen sharing due to the ongoing pandemic. Each participant was given four task prompts, three paired with a particular feature of the ArtBreeder platform, and the fourth where they were able to use all features concurrently. Participants were given 10 minutes to create a face for each prompt.

The task to design faces was selected for several reasons: faces are a rich creative domain and are the subject of many creative works, juxtapositions feature heavily in those works, ArtBreeder has a dedicated portrait model, and the domain does not require prior design expertise. We followed a mix between a semi-structured post-task interview and a concurrent think-aloud protocol: during each task participants were encouraged to describe what they were thinking and doing, as well as prompted if they remained silent. After each task participants were asked a few open-ended questions about their satisfaction with the result, creative self-efficacy, and the experience of using ArtBreeder.

The during- and post-task dialogues were transcribed, then coded using an inductive thematic analysis process (Braun and Clarke 2006) in NVivo 12[2]. The first round of coding sorted participant phrases into distinct themes, which were then organised into higher-order themes.

Results

272 participant quotes from our 11 participant interviews were sorted into 10 themes, which were in turn organ-

[2]https://lumivero.com/products/nvivo/

ised into three higher-order themes: Discovery and Open-endedness, Control and Intent, and Expressibility.

Table 1: Themes in "Discovery and Open-endedness"

Theme	Description
Design goals and strategies	Desired outcomes and intentions at the task and subtask levels, and the participant's means of achieving these desired outcomes
Search behaviour	Cognitive and creative process in finding and navigating tools and content that appropriately satisfy their desired outcome
Unpredictability and surprise	Unexpected and surprising behaviours, output, interactions, and performance of ArtBreeder systems and functions and to what degree they help or hinder the participant under different circumstances
The "black box"	Lack of understanding or intuition of the algorithms or functions in ArtBreeder, how the system works

Discovery and Open-Endedness

The Discovery and Open-endedness higher-order theme (see Table 1) represents behaviours and qualities relating to predictability, surprise, searching, novelty, exploration, non-fixation, associativity, fluid representations and re-representation. This higher-order theme involves the process of finding solutions, exploring the system and interacting with it, navigating through the index of user-generated content, and selecting the right candidate image from a neighbourhood of similarly appropriate images.

Table 2: Themes in "Control and Intent"

Theme	Description
First impressions	Immediate response to the ArtBreeder content, layout, community, functions, and interface - and any associations they make with other tools
User experience	Reflections on the interface, layout, and design choices and to what degree they interfere or support the participant's goals
Sense of control	Perceived level of influence over granular features and the final images produced

Control and intent

The Control and intent theme theme (see Table 2) represents the system's capacity to afford user control, the minimum predictability to achieve intent, the user experience of the ArtBreeder platform, and preconceptions towards the ArtBreeder website and community. Control and intent involves the specific properties users interact with directly or indirectly, specific features and components of ArtBreeder, and more generalisable beliefs about co-creative AI systems.

Table 3: Themes in "Expressibility"

Theme	Description
Creative self-efficacy	To what degree ArtBreeder has enabled them to express their creative intentions, or whether their cognitive experience of creativity has been augmented
Sense of authorship	Perceived ownership, level of input and directedness of the final output
Sense of collaboration	Perception of interaction and co-creation with an intelligent agent or with other users on ArtBreeder

Expressibility

Expressibility (see Table 3) represents the potential for participants to achieve creative reward and agency, impressions of authorship, and a sense of collaboration with both the ArtBreeder system and community. Participants expressed concerns over their control and authorship of artifacts made in a system that is largely unpredictable, difficult to understand, has varying levels of direct control, and the level of creative self-efficacy afforded by the system.

Discussion

Our case study highlighted both technical and design considerations related to how we might interact with intelligent creative design systems that possess some level of autonomy of intent and action. We discuss several of those considerations here, but overall our study emphasises the efficacy of MJTs as a model for co-creative systems research. Despite their simplicity, our tasks that involved a tension between concepts – one that is not easily facilitated by the system – required our designers to apply their human-level understanding and creativity. We observed users switching between reframing the problem and trying to solve it, reverting back to previous design states, and exploring serendipitous options afforded by the ArtBreeder tools. While our study was not controlled (in that we did not ask some users to perform tasks without essential conceptual juxtapositions), we can say from years of experience in co-creative systems and generative AI that we do not typically observe comparable levels of design-like thinking in other tasks of this simplicity. We believe accordingly that minimally juxtapository tasks are a promising approach for future co-creative AI studies.

In our ArtBreeder study, despite the simplicity of our tasks, the juxtaposition they required made them "creatively interesting" for our users, leading to a number of insights about co-creative systems. Users found the black box nature of ArtBreeder both compelling and frustrating, reflecting larger trends in artificial intelligence research: the need for explainable and interpretable models using techniques that make complex systems more transparent, communicative, and predictable (Llano et al. 2022; Zhu et al. 2018).

(a) Task 1: "A grizzled veteran with a heart of gold" by Participant 4

(b) Task 2: "A sweet senior with a wild side" by Participant 4

(c) "A detective with a music career side-hustle" by Participant 9

(d) "A zombie politician" by Participant 9

Figure 2: Sample participant-made images in each of our minimally-juxtapository tasks

Our users also experienced an expectation mismatch with ArtBreeder, as its resemblance to existing image manipulation tools made them apply their existing mental models from that domain. These mental models led them to expect much more direct control and fine-grained interactions, which would likely not have been the case if we had used a prompt-based or other more-abstract interaction paradigm. This is much broader than ArtBreeder or MJTs: as creative tools gain intelligence and autonomy and the line between tool and agent blurs, it is likely that existing expectations and mental models will be violated on a regular basis. This is not a bad thing, as new interaction modalities always require new interface paradigms, but it highlights the importance of the HCI work that must accompany the development of co-creative systems (Kantosalo et al. 2020).

The scope of the present study concerns generative co-creative systems; however, we believe that it is possible that MJTs could be of use in other computational design contexts. (Hayes et al. 2011) provides a retrospective summary of other forms of artificial intelligence in various design domains. It details research into reasoning systems that are model-based, knowledge-based and case-based; knowledge-representation and reasoning; generative design; and various research challenges and directions adjacent to the present study. In any design task where humans and intelligent agents collaborate, there is a potential for MJTs to be a useful framework for experimental design.

Even within generative co-creative systems, there is a wide variety of roles and modes of interaction. COFI, a framework developed on the study of 92 co-creative systems, identifies three fundamental interaction models: 1) generative agents that follow the user's directions, 2) mixed-initiative agents that work alongside users on a shared product, and 3) advisory agents that both generate and critique the user's creative product (Rezwana and Maher 2022). Artbreeder is an example of the former: it follows the user's directions, and the user must interpret the MJT's essential contradiction on their own. In a more mixed-initiative co-creative system like Reframer (Ibarrola, Bown, and Grace 2022) or the Drawing Apprentice (Davis et al. 2016), that creative responsibility (of somehow resolving the juxtaposition at the heart of the provided task) would be shared. This may stress the generative capacity of some systems, potentially leading to situations where the user must step in and take back control (as in our study), which could be a useful proxy for more-complex design tasks. A similar story may hold true for more advisory/critical agents, whose evaluative mechanisms may struggle in tasks with conflicting requirements.

Conclusion

We have conducted a user study designed around minimally juxtapository tasks (MJTs) to investigate the capacity for co-creative systems to support the solution to realistic conceptual design problems in the domain of concept art. The negotiation between conceptual tensions that occurs in early-stage creativity is critical, and surfacing it in user studies will hopefully attract additional attention to this underaddressed component. Our analysis shows that, even with these very simple tasks and a relatively simple (even outdated) co-creative system, a significant degree of nuance was achieved by our users in their design tasks. This suggests the potential of MJTs as a framing for experimental design in co-creative contexts. Our study also elicited a number of themes that show the challenge of designing future co-creative systems. Resolving some of these of these challenges will draw on the AI domain, namely explainability and shared goals and meaning, whereas others will draw on HCI and computational creativity, such as how to design interfaces and interactions that afford mixed-initiative collaboration.

References

Bentley, P. J., and Wakefield, J. P. 1997. Conceptual evolutionary design by a genetic algorithm. *Engineering design and automation* 3:119–132. Publisher: WILEY.

Bessette, J.; Fol Leymarie, F.; and W Smith, G. 2019. Trends and Anti-Trends in Techno-Art Scholarship: The Legacy of the Arts "Machine" Special Issues. In *Arts*, volume 8, 120. Multidisciplinary Digital Publishing Institute. Issue: 3.

Braun, V., and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3(2):77–101. Publisher: Taylor & Francis.

Davis, N.; Hsiao, C.-P.; Yashraj Singh, K.; Li, L.; and Magerko, B. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 196–207.

Davis, N. 2013. Human-computer co-creativity: Blending human and computational creativity. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 9, 9–12.

Di, X., and Patel, V. M. 2017. Face synthesis from visual attributes via sketch using conditional vaes and gans. *arXiv preprint arXiv:1801.00077*.

Dorst, K. 2015. *Frame innovation: Create new thinking by design*. MIT press.

Ford, C.; Bryan-Kinns, N.; Generative, A.; et al. 2022. Speculating on reflection and people's music co-creation with ai.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Hayes, C. C.; Goel, A. K.; Tumer, I. Y.; Agogino, A. M.; and Regli, W. C. 2011. Intelligent support for product design: Looking backward, looking forward. *Journal of Computing and Information Science in Engineering* 11(2).

Ibarrola, F.; Bown, O.; and Grace, K. 2022. Towards co-creative drawing based on contrastive language-image models. *coordinates* 10:2.

Kantosalo, A.; Ravikumar, P. T.; Grace, K.; and Takala, T. 2020. Modalities, Styles and Strategies: An Interaction Framework for Human-Computer Co-Creativity. In *ICCC*, 57–64.

Karimi, P.; Grace, K.; Davis, N.; and Maher, M. L. 2019. Creative sketching apprentice: Supporting conceptual shifts in sketch ideation. In *Design Computing and Cognition'18*, 721–738. Springer.

Lawton, T.; Ibarrola, F. J.; Ventura, D.; and Grace, K. 2023a. Drawing with reframer: Emergence and control in co-creative ai. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 264–277.

Lawton, T.; Ibarrola, F. J.; Ventura, D.; and Grace, K. 2023b. Drawing with reframer: Emergence and control in co-creative ai. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, 264–277. New York, NY, USA: Association for Computing Machinery.

Li, B.; Qi, X.; Lukasiewicz, T.; and Torr, P. H. 2020. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7880–7889.

Liapis, A.; Yannakakis, G. N.; Togelius, J.; et al. 2013. Sentient sketchbook: Computer-aided game level authoring. In *FDG*, 213–220.

Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2022. Explainable computational creativity. *arXiv preprint arXiv:2205.05682*.

Lourenço, N.; Assunção, F.; Maçãs, C.; and Machado, P. 2017. Evofashion: customising fashion through evolution. In *Computational Intelligence in Music, Sound, Art and Design: 6th International Conference, EvoMUSART 2017, Amsterdam, The Netherlands, April 19–21, 2017, Proceedings 6*, 176–189. Springer.

Maher, M. L.; Poon, J.; and Boulanger, S. 1996. Formalising design exploration as co-evolution. In *Advances in formal design methods for CAD*. Springer. 3–30.

Pahl, G.; Beitz, W.; Feldhusen, J.; and Grote, K. 2007. Engineering Design: A Systematic Approach Third Edition. *Berlin, Springer Science+ Business Media Deutschland GmbH, 2007. 632*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; and others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Rezwana, J., and Maher, M. L. 2022. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction*.

Wang, S.-Y.; Bau, D.; and Zhu, J.-Y. 2021. Sketch your own gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14050–14060.

Wiltschnig, S.; Christensen, B. T.; and Ball, L. J. 2013. Collaborative problem–solution co-evolution in creative design. *Design Studies* 34(5):515–542.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhu, J.; Liapis, A.; Risi, S.; Bidarra, R.; and Youngblood, G. M. 2018. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE conference on computational intelligence and games (CIG)*, 1–8. IEEE.

Solving and Generating NPR Sunday Puzzles with Large Language Models

Jingmiao Zhao and Carolyn Jane Anderson
Computer Science Department
Wellesley College
Wellesley, MA 02482 USA
carolyn.anderson@wellesley.edu

Abstract

We explore the ability of large language models to solve and generate puzzles from the NPR Sunday Puzzle game show using PUZZLEQA, a dataset comprising 15 years of on-air puzzles. We evaluate four large language models using PUZZLEQA, in both multiple choice and free response formats, and explore two prompt engineering techniques to improve free response performance: chain-of-thought reasoning and prompt summarization. We find that state-of-the-art large language models can solve many PUZZLEQA puzzles: the best model, GPT-3.5, achieves 50.2% loose accuracy. However, in our few-shot puzzle generation experiment, we find no evidence that models can generate puzzles: GPT-3.5 generates puzzles with answers that do not conform to the generated rules. Puzzle generation remains a challenging task for future work.

Introduction

Puzzles and games have long been used to benchmark progress in AI. We continue this tradition by exploring the ability of large language models (LLMs) to solve word puzzles from the NPR Sunday Puzzle on-air game show. Recent advances have led to new techniques for using general-purpose text generation models to solve a variety of tasks. In few-shot learning, a model is prompted with a handful of examples and asked to generate a solution. In prompt engineering, the input to the model is manipulated in order to improve the model's performance on a task. These techniques have led to surprisingly good performance by LLMs on novel tasks, without any further training of the model.

In this paper, we explore whether few-shot learning and prompt engineering can allow LLMs to solve questions from the NPR Sunday Puzzle game show, which combines information retrieval, wordplay, and pattern recognition. We introduce PUZZLEQA, consisting of puzzle descriptions, questions, and answers for 558 puzzles, and use it to benchmark four state-of-the-art LLMs. We explore prompt engineering techniques, but find that they have little impact on performance. We also explore whether models can generate new puzzles and find that this remains a challenging task. Although the best model, GPT-3.5, is capable of solving 50.2% of the puzzles, it cannot generate playable games.

Puzzle Description: Today's puzzle involves "consonyms," which are words that have the same consonants in the same order but with different vowels. Every answer is the name of a country.
Question: MINGLE
Answer: MONGOLIA

Figure 1: NPR Sunday Puzzle from March 12, 2023

Benchmarking AI through Games

Our work continues the tradition of evaluating AI progress through puzzles and games (Ferrucci 2012; Rodriguez et al. 2021; Rozner, Potts, and Mahowald 2021; Sobieszek and Price 2022). Contemporary LLMs have demonstrated strong performance on a wide variety of language tasks, including question-answering. However, the extent of their ability to generalize patterns and to solve wordplay puzzles is underexplored.

The NPR Sunday Puzzle game show represents a particularly interesting genre of puzzle to explore because it synthesizes a variety of skills: information retrieval; rhyming, anagram-solving, and other wordplay; and pattern recognition. Figure 1 shows an example of a puzzle, which involves knowledge of country names and wordplay. Despite the complexity of some NPR Sunday Puzzle games, compared to other question-answering games used to benchmark LLMs, such as Jeopardy! and Quiz Bowl, they are targeted towards a broader audience and require less specialized knowledge.

Dataset

We present PUZZLEQA, a dataset of 558 NPR Sunday Puzzle games from 2007-2021. During this period, a group of fans ran a mailing list, NPR Puzzle Synopsis, that distributed questions and answers for each week's puzzle.[1] We obtained the puzzle explanations from the NPR website,[2] and extracted the answers from the mailing list, using GPT-J to aid in preprocessing the data. We also classified the puz-

[1] https://groups.google.com/g/nprpuzzle
[2] https://www.npr.org/series/4473090/sunday-puzzle.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

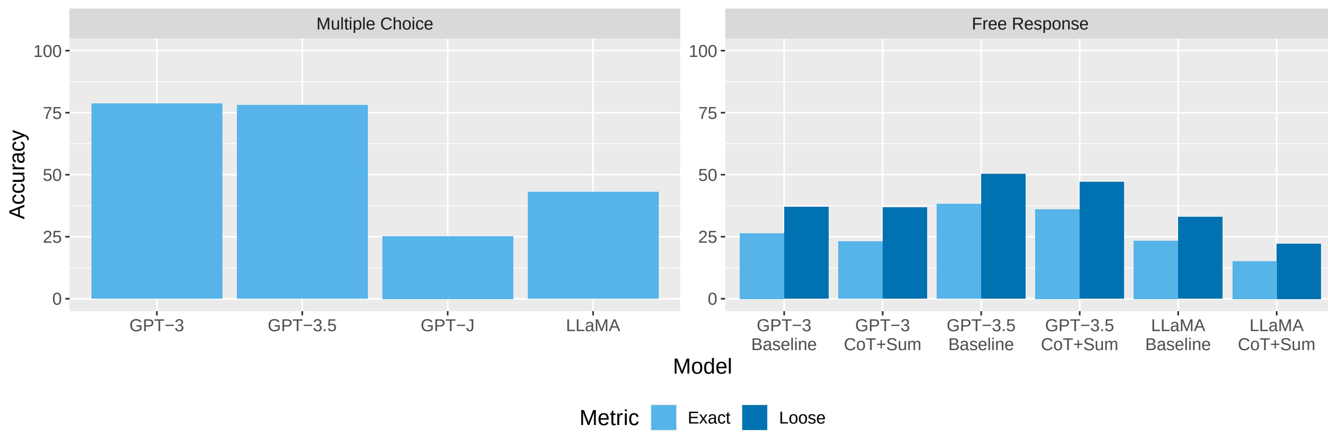


Figure 2: Results on full PUZZLEQA dataset, by model, prompting technique, and format.

zles into 11 different categories. The dataset, preprocessing tools, and analysis scripts will be released publicly.[3]

Model Selection

We explored two publicly available LLMs, GPT-J (Wang and Komatsuzaki 2021) and LLaMA (Touvron et al. 2023), and two proprietary OpenAI LLMs: GPT-3 Davinci and GPT-3.5 (Brown et al. 2020). The amount of randomness in each of these model's generations can be manipulated via the temperature hyperparameter, where a high temperature means more randomness. After exploring temperature settings of 0.75, 0.5, 0.25, and 0.1, we found that temperature = 0.1 was optimal.

Multiple Choice Experiments

As an easier benchmark, we constructed a multiple-choice version of the PUZZLEQA dataset. For each problem, we randomly selected three answers to other questions from the same puzzle to present alongside the correct answer.

Answer only baseline In multiple choice tasks, there can be biases towards or against certain question options, even in the absence of the question. To obtain an accurate baseline, we measure how often the model selects the correct answer when it is not given the question. An unbiased set of answer options would result in at-chance performance (25%). We refer to this task as the answer-only baseline. We find that the model selects the correct answer 21% of the time when it is not given the question, suggesting that there is no significant bias towards the correct answer from the answer options alone.

Results Figure 2 shows the performance of each model on the multiple choice task. The smallest model, GPT-J, does not perform better than chance on this simplified task. As a result, we exclude it from the rest of our experiments. The other publicly available model, LLaMA, performs well above chance, showing that it is able to correctly identify responses for many problems. GPT-3 and GPT-3.5 both perform well on this task, solving 78% percent of problems.

[3]https://github.com/Wellesley-EASEL-lab/PuzzleQA

Free Response Experiments

We perform two sets of free response experiments. To explore various prompt engineering techniques, we first create a subset of our data balanced by question type. We then compare the best prompting technique against a baseline on the full dataset. In all experiments, a few-shot paradigm is used: the model is given two examples of solved questions from the same puzzle (following the same game rules) and asked to solve a third.

Evaluation Metrics

Free response question-answering is difficult to evaluate, since a correct answer may be phrased in various ways. We use two conservative metrics for evaluating performance.

Exact Matching: the response is correct if it exactly matches the gold solution.

Loose Matching: the response is correct if it is contained or contains the gold solution, after removing non-alphabetical characters and lowercasing both strings.

Exploring Prompt Engineering Techniques

We subsample our dataset in order to evaluate the impact of various prompt engineering techniques. 10 questions from each of our 11 categories were randomly sampled for the subset, for a total of 110 items. We explore two prompt engineering techniques: summarization and chain-of-thought reasoning.

Summarization One potential challenge for the model in solving the PUZZLEQA puzzles is that the games are described informally. We hypothesize that the lack of consistency in puzzle wording might hinder the model. We experiment with using GPT-3.5 to summarize the puzzle description to a more consistent format (Figure 3).

Summarize the following: In the on-air puzzle, you are given the word and must drop two letters so that the remaining letters, in order, spell a color or shade.

Figure 3: Summarization prompt to summarized explanations of the rules of the puzzle

Puzzle Description: In the on-air puzzle, you are given the word and must drop two letters so that the remaining letters, in order, spell a color or shade.
Question: blouse
Answer: blue
Please explain this answer.

Figure 4: Chain-of-thought prompt to elicit explanations

Chain-of-thought Reasoning Prompting models to explain their reasoning before generating an answer has been shown to improve model performance on other tasks (Wei et al. 2023). This is known as *chain-of-thought prompting*.

One limitation of this approach is that humans must write explanations to provide as examples to the model. We automate the process by using the model to generate explanations for rule-question-answer triplets. We then use the generated explanations as input to the chain-of-thought prompting experiment. Figure 4 shows an example prompt used to gather model explanations. GPT-3's generated explanation was *The word "blouse" can have two letters dropped to spell the color "blue"*. This explanation was then added to the example to use in few-shot prompting.

Prompt Engineering Results

In our small-scale experiment, we found that both summarization and chain-of-thought prompting improved performance. Figure 5 shows GPT-3 results for each technique.

Free Response Results

We select the best-performing prompt engineering technique to compare against a baseline prompt on the full PUZZLEQA dataset. Our small-scale experiments suggested that both summarization and chain-of-thought prompting improve performance. We compare this model to a baseline few-shot learning model.

Figure 2 shows the free response results for the full PUZZLEQA dataset. Although chain-of-thought reasoning and summarization improved model performance in our small-scale experiment, this did not replicate for the entire dataset. The baseline GPT-3.5 model performs best, solving 50.2% of the puzzles. We note that performance is very sensitive to prompt wording: when we rephrase the chain-of-thought prompt to ask for the "answer and reasoning" rather than the "reasoning and answer," performance drops substantially.

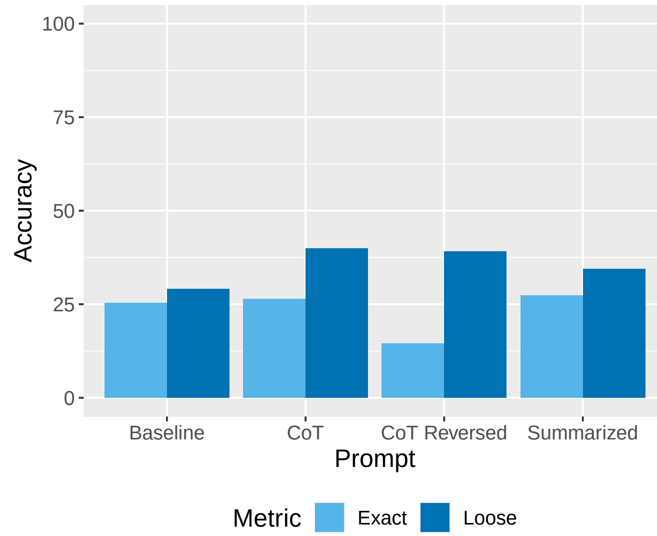


Figure 5: GPT-3 results for prompt engineering experiment

Discussion Overall, the best model, GPT-3.5, performs fairly well on the PUZZLEQA dataset. We observe a large gap in performance between the LLaMA model, which is available for academic research, and the proprietary OpenAI models, illustrating the need for better public LLMs. Surprisingly, we find that the prompt engineering techniques we explored did not improve model performance.

Although chain-of-thought prompting did not improve performance, we feel that it still has some benefits. The chain-of-thought-prompted model: when prompted this way, GPT-3.5 produces "N/A" 87 times, compared to only 3 refusals to provide an answer in the baseline version. In some cases, the puzzle is faulty due to webscraping errors. When we manually examined the explanations generated by the model, most are consistent with the answer. Thus, chain-of-thought prompting may decrease overconfidence in models while providing a window into the model's decisions.

Game Generation

Although our exploration of popular prompt engineering techniques was not fruitful, we nonetheless found that state-of-the-art LLMs are capable of solving many of the NPR puzzles. In this section, we explore whether LLMs are also capable of generating puzzles for humans to play. We explore puzzle generation with the LLM that achieved the highest performance on the free response task, GPT-3.5.

Prompt Design

We construct a few-shot puzzle generation dataset using our balanced 110 question subset. In each prompt, we provide the model with five examples of puzzles, presented as explanation-question-answer triplets (Figure 6), for a total of 22 prompts. We then evaluate each generated puzzle by asking the model to generate an answer to the question.

Explanation	Question	Response	Gold
I'll give you a word that can be split into two smaller words. The first word is the name of a U.S. state, and the second word is a type of animal. For example, given "MontanaLion," the answer would be "Montana, Lion."	DelawareHawk	Delaware, Hawk	Delaware, Hawk
Every answer is a well-known movie title with one or more letters from the title replaced with a number. For example, if the clue is "Th3 Matrix," the answer would be "The Matrix."	F1ght Club	Fight Club	Fight Club

Table 1: Game that satisfies both consistency and conformity, but are trivial

> You are given several examples of the game, with each game including a prompt, question, and answer.
> *5 examples given as:*
> *Explanation:*
> *Question:*
> *Answer:*
> Please generate a new game with a prompt, question, and answer in the same format.

Figure 6: Game generation prompt

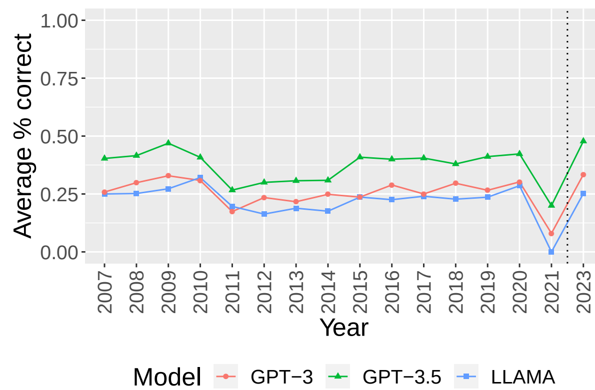


Figure 7: Exact accuracy by puzzle air date

Evaluation Metrics

We use two metrics to evaluate the generated games:

Consistency: can the model solve its own puzzle? We provide the generated explanation and question to GPT-3.5 and generate an answer. If the answers match, the puzzle is consistent.

Conformity: of the questions that are consistent, how many have answers that conform to the rules in the explanation? We assess this manually.

Results

Of the 22 games generated by GPT-3.5, it answers 17 questions consistently. However, just 6 of the questions conform with the explanation provided. In addition, the conforming games are trivial to solve (Table 1). Thus, though LLMs succeed in playing the NPR Sunday Puzzle, we find no evidence that they can generate new puzzles for human players.

Limitations and Future Work

Our experiments with PUZZLEQA show that current LLMs are capable of solving, but not creating, NPR Sunday Puzzle questions. However, our results come with a number of caveats. First, since the training data for GPT-3, GPT-3.5, and LLaMA is not publicly available, we cannot measure whether models have been trained on problems within our dataset. To investigate potential training/test overlap, we manually constructed a test set of questions from 2023, which is more recent than the models' training data cutoff dates (GPT-3: 2019; GPT-3.5: 2021; LLaMA: 2022). We find that model performance on this small (n=116) subset is on par with the full dataset (Figure 7). In general, although performance varies by year, there is no clear trend.[4]

Our methodology could also be refined in a number of ways. Our webscraping techniques failed to capture some questions, which could be added to our dataset. Our loose accuracy metric is a conservative measure of model capability, since it may fail to identify some valid answers. Finally, future work could incorporate a rating of question difficulty by identifying from the game transcript whether the human player succeeded or failed in answering the question.

Conclusion

Using data from the NPR Sunday Puzzle game show, we explore the ability of contemporary large language models to solve and generate word puzzles. We show that PUZZLEQA is a challenging benchmark for LLMs: although GPT-3.5 solves 50.2% of the problems in the free response task, information about its training data is not public, and the best publicly available model achieves only 33%.

The fact that the prompt engineering techniques we explored failed to improve performance is puzzling, given promising results from chain-of-thought prompting reported for similar tasks (Wei et al. 2023). However, we argue that chain-of-thought reasoning is still helpful for explainability.

Our game generation results show that being able to generate NPR Sunday Puzzle-style games is beyond the capabilities of current LLMs, even if they are capable of solving

[4]We note that puzzle types and topics may vary over time; an in-depth analysis of the puzzle content is one area for future work.

them. Future work could explore fine-tuning a model on our dataset rather than using few-shot learning. We hope that the PUZZLEQA dataset will aid future work in this area.

Author Contributions

All authors contributed to the writing of this paper.

Acknowledgements

We would like to thank our anonymous ICCC reviewers for their helpful feedback. We also extend our thanks to Eni Mustafaraj and Brian Tjaden for their comments on this work at various stages in its development, and to Arjun Guha for the initial idea of exploring the NPR Sunday Puzzle game.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.

Ferrucci, D. A. 2012. Introduction to "This is Watson". *IBM Journal of Research and Development* 56(3.4):1:1–1:15.

Rodriguez, P.; Feng, S.; Iyyer, M.; He, H.; and Boyd-Graber, J. 2021. Quizbowl: The case for incremental question answering.

Rozner, J.; Potts, C.; and Mahowald, K. 2021. Decrypting Cryptic Crosswords: Semantically Complex Wordplay Puzzles as a Target for NLP. arXiv:2104.08620.

Sobieszek, A., and Price, T. 2022. Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines* 32(2):341–364.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and efficient foundation language models.

Wang, B., and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Constraints as Catalysts: A (De)Construction of Codenames as a Creative Task

Brad Spendlove and Dan Ventura
Computer Science Department
Brigham Young University
Provo, UT 84602 USA
brad.spendlove@byu.edu, ventura@cs.byu.edu

Abstract

Constraints are a common feature of creative domains, and the presence of constraints often facilitates creative outcomes. We examine the relationship between constraints and creativity by examining the competitive language game Codenames. We characterize the game space of the Codenames spymaster role by describing a set of successive constraints that give rise to the game. This constraint-centric characterization both demonstrates that the game is successfully designed to allow for creative play and also serves as a basis for a computational analysis of the spymaster task. We consider some of the implications of this characterization generally, and how we can think about both the game and the abstract principles it instantiates from the standpoint of a computationally creative system for playing it and what we might learn about creative search by building such a system.

Introduction

The concepts of constraint and creativity seem to enjoy an intimate relationship: creativity is necessary when constraints are present and constraints are necessary for creativity to be present. In what follows, we illustrate this in the context of a competitive language game, presenting a series of thought experiments while "designing" the game and examining a critical constraint "phase transition" that appears crucial to the game's success both as a game and as a creative exercise. We then consider the task imposed by the game from a computationally creative standpoint.

Competitive language games have recently been suggested as an interesting domain for creativity because they offer a proxy measure for (successful) creativity in the form of winning the game (Spendlove and Ventura 2022). Codenames (Chvátil 2015) is a well-known example of the genre, and we will assume a familiarity with it as we proceed, as it provides an interesting case study for our arguments.

To begin with, we note that constraints play a critical role in making the game an interesting/challenging/fun game, and we will demonstrate that momentarily. First, though, we note the existence of a meta-level creative task, presented by the game,[1] of figuring out the best general strategy given the constraints. The rules of the game (constraints) are designed to force this meta-level strategy to be one that requires new (base-level) creativity to solve each new game. We proceed by iteratively building up various constraints to illustrate the "emergence" of both the game of Codenames itself as well as (the need for) creativity in the context of the game. In a very real sense, these seem to exhibit a primal/dual relationship.

At an abstract level, the unconstrained task is to reveal 9 locations out of 25. That means there are $\binom{25}{9}$ different target states, with each instance of the game requiring the communication of one of them. Without additional rules forbidding this, one can easily accomplish the task by simply pointing to the 9 target locations. This is a solution strategy that wins the game in a single turn and is target-state agnostic (that is, it works equally well for any target state). While there are likely many ways of accomplishing this task, anyone first exposed to the game task is likely to immediately "invent" this solution, which requires only that the players can see and the spymaster can point. It is not at all surprising, and if it is employed, the game is no game at all.

However, we can introduce our first constraint by allowing only verbal communication, because this is a language game. What strategy might then be employed instead? We can invent a simple indexing scheme for the 25 locations and then give a sequence of 9 indices to communicate the targets. This strategy requires that the players understand the indexing system (they have to get together in advance to communicate it, or hope it is "self-evident" enough that it can be picked up on the fly). This is again a solution strategy that wins the game in a single turn and is again target-state agnostic.

We can introduce a further constraint by allowing the spymaster to communicate only a single clue word per turn.[2] A somewhat clever strategy involves constructing a mapping from the $\binom{25}{9}$ possible positions to the integers. Then, the clue word is just the appropriate integer.[3] To construct such a mapping, consider the board as an element of $\mathbb{B}^{25}$. Then the

[1] It is possible that many other or even all other creative tasks also present this same meta-task, though we don't explore that claim here.

[2] This ignores the additionally allowed clue number, but for the current discussion this is unimportant.

[3] For the sake of argument, we consider any integer a single "word".

integer for each board position is given by the binary number that has 1s in the 9 target positions and 0s elsewhere. This requires that the players understand binary numbers and how to convert between base 10 and base 2 representations. This is again a solution strategy that wins the game in a single turn and is again target-state agnostic.

To impose further constraint, we can disallow the binary code strategy by being a stickler about word count or by disallowing numbers.[4] A strategy for meeting this additional level of constraint involves the invention of a diabolical code that maps an ordered list of $\binom{25}{9}$ English words to the $\binom{25}{9}$ integers above, which works as follows. Associate the ith position in the word list with the ith 25-bit binary number that contains exactly 9 1s. This requires that the players have the same ordered list of words (which must be communicated in advance). In addition, if visual aids are disallowed, the players must be able to commit the list to memory. This is yet again a solution strategy that wins the game in a single turn and is also target-state agnostic. It also is beyond human capability.[5] Note that at each additional constraint level, we have produced a winning meta-level strategy: they all immediately win the game for any target state.

The constraints have become restrictive enough now to suggest a couple of potentially interesting Codename-specific questions:

- What is the best trick like this diabolical code that is not beyond human capability?[6]
- Does there exist such a strategy that wins in a single turn and is target-state agnostic?
- If not, does there exist a target-state agnostic strategy of any turn length?
- If so, what is the minimum number of turns required for the optimal target-state agnostic "cheat code"?

To constrain the task even further, we can disallow such cheat codes by requiring that the clue word "must be about the meaning of the [target] words".[7] Given this, we have reached the actual instantiation of the game of Codenames. And, it is interesting to note that it is now unclear that there exists any target-state-agnostic strategy that wins the game in a single turn, even given superhuman ability.[8] A corollary to this is that the game's rules have been designed in just such a way as to make the answer to this meta-level strategy question be that the best strategy now is game-specific.

[4]Numbers are allowed to be used in Codenames in some ways but not in others.

[5]And there's another difficulty—according to Oxford, there are currently $171,476 \approx \binom{25}{6}$ words in use in the English language, which means that there are roughly an order of magnitude fewer (current) English words than we require to implement our diabolical code solution.

[6]That is, what is the best "cheat" we could reasonably operationalize as human players?

[7]cf. Codenames rules.

[8]That is to say, we have not been able to invent such a strategy and therefore leave it as future work and/or an exercise for the creative reader to either produce such a strategy or prove it is not possible to do so.

Assuming this is the case, the natural follow-on question is, what is the game-specific strategy for winning as quickly as possible:

- Is there a game-specific, guaranteed, one-move strategy?

The constraints of the game appear to have transmuted the task, which now requires a new level of game-specific creativity—barring an epiphany that allows us to communicate positions directly, we instead are reduced to trying to communicate semantic relationships among words, by giving a single clue-word. This transmuted task can be represented as the discovery of meaningful relationships amongst a set of words. These relationships are dependent on the persons involved, their experiences, their knowledge and any shared knowledge/experience, and will evolve over time; as a result, there is no correct solution to the task (a hallmark of creativity "problems"), and such relationship artifacts can be novel, valuable, and surprising (more creativity hallmarks).

What is the key to the apparent "phase transition" from a less-constrained game that admits a meta-solution to the more-constrained game that (apparently) admits only game-specific solutions? For game variants with fewer constraints, all target states are in some important sense indistinguishable, so it is possible to find a general solution (which can still require creativity, of course). By contrast, as the constraints are increased to the point where semantic (or other types of) relationships between words become important, these relationship artifacts are no longer indistinguishable, so finding a general solution is nontrivial at best (and may likely be impossible).

Perhaps there is an argument to be made here that the genius of the game designer is in imposing constraints that disallow "cheating"/boring target-state-agnostic meta-solutions, while still maintaining enough flexibility to admit fun/creative target-state-specific solutions.

Clue Graphlets

The artifact created by a Codenames spymaster is more than just a word w (and number k); it is a graphlet of connections between words. The clue word w, drawn from the set of all English words, is the center node in the graphlet. The k word cards the spymaster intends to relate to the clue word are each connected by an edge to the center node. The center clue word and the number of connections k are given to the spymaster's teammates, and their task is to guess which word cards $\{c_1 \ldots c_k\}$ fill in the graphlet. Figure 1 shows an illustration of this structure for $k = 4$.

For example, if the spymaster's team's word cards include "Plane" and "Ambulance", a potential clue word that relates to both could be "Vehicle." We can formulate this as a graphlet with "Vehicle" in the center, an edge to "Ambulance" and an edge to "Plane".

There are many ways that two words can be related. For the purposes of playing Codenames, however, we are only concerned with whether a potential clue word will direct the guessers to a given word card or not. Despite the many different forms this relationship could take, in practice it is usually intuitive for humans to determine. For example, "Vehicle" could serve as a clue for "Ambulance", but "Sky" would

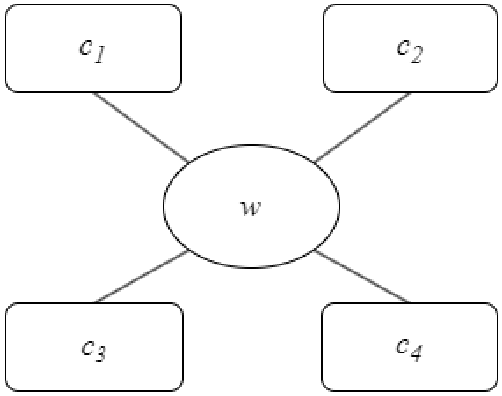


Figure 1: The structure of a Codenames clue graphlet with $k = 4$, consisting of a potential clue w and four word cards $c_1 ... c_4$.

likely not.[9] We can abstract this by considering a function $rel(w, c)$ that takes two words w and c, and returns `True` if w relates to c in this way or `False` otherwise.[10] This function can be used to evaluate the edges of a given graphlet. Call the potential clue word w and the set of word cards under consideration C, with $|C| = k$. If $rel(w, c_i)\ \forall c_i \in C$, then w is a good clue word for C.

The spymaster's job is to search through the set of English words for one that relates so well to k target words that the guesser's job is easy. Although the graphlet is the spymaster's creative artifact, only the clue word w and clue cardinality k are given to the guesser. The better the clue word, the more obvious it makes the graphlet's relationships. Note that this constraining of communication between the spymaster and the guesser(s) suggests an interesting alternative interpretation/understanding of the game—it is a game about a type of co-creativity—the spymaster creates an interesting/useful graphlet and then attempts to help the guesser discover that same graphlet by giving a hint. Viewed through the lens of creativity, the game requires creativity from both the spymaster and the guesser, with the spymaster first being creative and then guiding the guesser(s) to be creative in the same way.

Spymaster Task Search Space

To analyze the computational difficulty of the spymaster's task, we will first reason about the number of potential graphlets that the spymaster must search for a clue pair (w, k). For one move of the game, the board will have at most 9 word cards belonging to the spymaster's team. The spymaster may choose any number of those cards to incorporate into their chosen graphlet. An upper bound on the total number of unique configurations of word cards included (or not) in the graphlet is therefore $(\binom{9}{0} + \binom{9}{1} + \cdots + \binom{9}{9}) = 2^9 = 512$.

For each of these graphlets, the spymaster must consider English words to fill in the center node. Given an estimate of $170,000$ English words in current use, that puts the total number of graphlets to be searched at $2^9 \times 170000 \approx 8.70e7$. These graphlets, however, only consider the relationships between the potential clue word and our team's word cards. If the potential clue also relates to one of the other team's words, a neutral word, or the assassin word, then that graphlet is not a good candidate to be chosen as the spymaster's clue that round. We observe that a single such undesirable relationship disqualifies a graphlet, so any graphlet needs only to be checked individually against the 16 disqualifying word cards (those not belonging to the spymaster's team), increasing the size of the search space to $2^9 \times 16 \times 170000 \approx 1.39e9$.

We can significantly reduce the number of graphlets that must be considered by only searching for graphlets with up to four word cards. According to the Codenames rulebook, choosing a successful clue that relates to four word cards is a difficult accomplishment. A system that could consistently create clues (w, k) with $k = 4$ would perform at a superhuman level. We can also exclude the trivial case of a graphlet with zero connections, as it is irrelevant to the game. This results in a reduction of the size of the search space to $(\binom{9}{1} + \binom{9}{2} + \binom{9}{3} + \binom{9}{4}) \times 16 \times 170000 = 6.94e8$ graphlets.

We can further reduce the size of this search space by observing that graphlet edges must be considered one at a time. Thus, to evaluate a graphlet with $C = \{c_1, c_2, c_3, c_4\}$, the agent must first evaluate the graphlets with $C = \{c_1\}$, $C = \{c_1, c_2\}$, and $C = \{c_1, c_2, c_3\}$. By caching those calculations, the agent does not need to (explicitly) consider graphlets of $k < 4$ separately. This leaves the agent with $\binom{9}{4} \times 16 \times 170000 = 3.43e8$ potential graphlets to search.

Finally, note that a rough estimate of a college-educated person's vocabulary is $30,000$ words (D'Anna, Zechmeister, and Hall 1991). Substituting that for $170,000$ in our calculations results in a search space containing $6.05e7$ graphlets. Therefore, reasonable estimates of an agent's vocabulary do not change the search space by more than an order of magnitude.

Given a vocabulary size and maximum graphlet size (e.g. $30,000$ and $k = 4$ respectively above), the number of graphlets likely cannot be further reduced *a priori*.[11] Further reduction of the search space (at compute time) would require employing search heuristics. We observe that even without employing such heuristics, the number of graphlets through with the spymaster must search is of a magnitude that is potentially computationally tractable. The determin-

[9]Of course, some non-obvious connection could be drawn between "Sky" and "Ambulance", but this would require that the spymaster and guesser both are aware of that relationship, which raises theory of mind questions.

[10]In practice, this function could return a more nuanced value, such as a real number in the range [0,1]. However, we can simplify this by assuming a threshold to convert the real into a Boolean.

[11]Another approach to refining the bound could be characterizing the connectivity of the specific word cards included in the Codenames deck. Cursory examination suggests that the word cards are especially evocative or easy to draw connections between. More rigorous data analysis may be able to identify relevant characteristics of those cards.

ing factor is not the size of the search space of graphlets, but the cost of evaluating them.

The Dual Nature of Human and Computer Spymasters

When designing creative computer systems, we naturally turn toward human performance at the creative task as a gold standard. In the case of the Codenames spymaster, it is clear that human players can execute the spymaster task (and enjoy doing so!) A primary tool in accomplishing this is our powerful language faculty. Although such abilities are by no means easy to implement computationally, we can use them as the basis for computational analysis of the spymaster task.

Whether an agent has a fast or powerful method for determining the relationships between words or not, the computational task is the same. All graphlets in the reduced set described previously are candidates for clues. Therefore, it is instructional to compare search strategies for each agent over the set of graphlets given a function rel that determines whether a relevant relationship exists between two words.

The computational cost for searching for a clue graphlet is the product of the time it takes to search the space of all graphlets and the time it takes to execute rel on each edge in those graphlets. We can therefore reason about the costs of four computational tasks: human search, human rel, computer search, and computer rel.

Human language faculties include the storage of complex networks of semantically related concepts (Collins and Loftus 1975). Given both an understanding of human language faculties and observations of successful human Codenames play, it can be inferred that humans can execute rel with significantly higher speed and accuracy compared to a computational implementation.

Humans and computers employ different search strategies, each drawing on their own strengths (He, Mao, and Boyd-Graber 2022). While it is difficult to reason about the exact human search strategy for the spymaster task, we can assume from many existing examples that the computer search will be faster. Computational search strategies are well-understood, and selecting the best for the task gives us a lower bound on computer search speed.

This leaves us in the curiously complementary situation that *the human spymaster has a fixed, efficient rel function and the computer spymaster has a fixed, efficient search function*.

The skill and creativity exhibited by the human playing the spymaster reside primarily in their ability to effectively navigate the imposingly large search space. The better the human can do that, the better they will perform at the overall task.

This aligns with intuition and observation of human Codenames play. The human spymaster brings a practically immutable set of semantic relationships to the game and can exercise different search strategies in an attempt to maximize performance. As this work primarily concerns computational creativity, we defer further exploration of human spymaster strategies to future work.

With a computer agent's relatively low search cost, the gap between computer and human spymaster performance comes down to the cost of the computer's rel function. Methods for designing efficient implementations of this and related functions are open research questions. Computer models of semantic meaning have been the subject of ongoing study in the fields of natural language processing and machine learning (Otter, Medina, and Kalita 2020).

One approach to isolating and analyzing this function is employing ideal module prototyping (Spendlove and Ventura 2020). This paradigm sees creative system designers replacing flawed computer task modules with human-delegated versions. By so doing, the system's architecture can be tested without the confounding factors of inefficient or incorrect task modules. Once the module-agnostic design is validated, effort can be expended to improve the flawed module with the assurance that any low-quality output is not due to other factors.

Discussion

We have demonstrated that reasoning about Codenames as a set of constraints facilitates a thorough characterization of the game and its creative tasks. It also allows us to rule out any potentially pathological play patterns. The constraints implicit in the game's design delineate a search space of clue graphlets that we have explored in some detail.

The designs of other games may also be deconstructed in a similar way to allow for more explicit characterization of the games' search spaces. Of course, for some types of games, such as abstract strategy, this analysis may be trivial and unnecessary. Language games, however, are a popular genre of game that inherit the complexity and open-endedness of human language. Language games may be a source of many well-defined creative tasks that could be excellent candidates for computational creativity research. Our constraint-centric analysis can serve as a template for analysis of such games and their creative tasks. Future work may generalize this hierarchical constraint characterization to creative tasks in general.

Our analysis of Codenames highlights the relationship between rules and creativity, demonstrating how the introduction of certain constraints can act as an intentional catalyst for creativity. We observe that the addition of specific constraints to expansive or mundane game spaces can unlock the potential for creative gameplay.

We have demonstrated how adding constraints to a trivial word identification game transforms it into the intriguing and creative game of Codenames. These constraints act as focal points that channel players' attention, encouraging them to explore, experiment, and discover creative solutions. Constraints establish boundaries, rules, or objectives that guide players' actions, transforming an otherwise unmanageable or uninspiring space into a captivating and intellectually stimulating environment. An interesting question for future work is if the abstract notion of hierarchical characterization of constraint generalizes to other specific (language) games and/or example tasks.

Another angle for future work could address the possibility that the addition of specific constraints may induce a phase transition from games (or other types of tasks/domains) that do not allow for creativity and those that do. The concept of a phase transition, commonly encountered in physics and other complex systems, refers to a qualitative change in the behavior of a system as the result of external factors or internal conditions. In the context of game design, we have shown an example of how the introduction of specific constraints can lead to a significant shift in the creative dynamics within the game. The obvious follow-on question is whether there is a general principle that elucidates the relationship between constraints and creativity.

References

Chvátil, V. 2015. *Codenames*. Kladno, Czech Republic: Czech Games Edition. Board Game.

Collins, A. M., and Loftus, E. F. 1975. A spreading-activation theory of semantic processing. *Psychological Review* 82(6):407.

D'Anna, C. A.; Zechmeister, E. B.; and Hall, J. W. 1991. Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior* 23(1):109–122.

He, W.; Mao, A.; and Boyd-Graber, J. 2022. Cheater's bowl: Human vs. computer search strategies for open-domain QA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3627–3639.

Otter, D. W.; Medina, J. R.; and Kalita, J. K. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32(2):604–624.

Spendlove, B., and Ventura, D. 2020. Humans in the black box: A new paradigm for evaluating the design of creative systems. In *Proceedings of the 11th International Conference on Computational Creativity*, 311–318.

Spendlove, B., and Ventura, D. 2022. Competitive language games as creative tasks with well-defined goals. In *Proceedings of the 13th International Conference on Computational Creativity*, 291–299.

Pushing GPT's Creativity to Its Limits: Alternative Uses and Torrance Tests

Fabricio Goes[1], Marco Volpe[1], Piotr Sawicki[2], Marek Grześ[2], Jacob Watson[3]
[1] Computing and Mathematical Sciences Department, University of Leicester, UK
[2] School of Computing, University of Kent, Canterbury, UK
[3] National Space Centre Creative, Leicester, UK
fabricio.goes@leicester.ac.uk, marco.volpe@leicester.ac.uk, p.sawicki@kent.ac.uk,
m.grzes@kent.ac.uk, jacobw@nsccreative.com

Abstract

In this paper, we investigate the potential of Large Language Models (LLMs), specifically GPT-4, to improve their creative responses in well-known creativity tests, such as Guilford's Alternative Uses Test (AUT) and an adapted version of the Torrance Test of Creative Thinking (TTCT) visual completion tests. We exploit GPT-4's self-improving ability by using a sequence of forceful interactive prompts in a multi-step conversation, aiming to accelerate the convergence process towards more creative responses. Our contributions include an automated approach to enhance GPT's responses in the AUT and TTCT visual completion test and a series of prompts to generate and evaluate GPT's responses in these tests. Our results show that the creativity of GPT's responses can be improved through the use of forceful prompts. This paper opens up possibilities for future research on different sets of prompts to further improve the creativity convergence of LLM-generated responses and the application of similar interactive processes to tasks involving other cognitive skills.

Introduction

Creativity tests are crucial instruments for evaluating human creative skills. One notable instance is Guilford's Alternative Uses Test (AUT) (Guilford 1967), which gauges divergent thinking by asking individuals to come up with as many different uses as they can for everyday items, such as a fork or a paperclip. Another commonly employed creativity evaluation is the Torrance Test of Creative Thinking (TTCT) (Torrance 1966), which consists of both verbal and visual tasks. The verbal aspect requires participants to produce ideas, hypotheses, or resolutions in response to open-ended questions, while the visual aspect entails completing partially drawn shapes or figures in an innovative and imaginative way.

In recent years, Large Language Models (LLMs), such as GPT (Generative Pre-trained Transformer) from OpenAI, have demonstrated impressive creative capabilities comparable to humans in generating jokes, poetry and other tasks (Toplyn 2022; Goes et al. 2022; Sawicki et al. 2023; OpenAI 2023; Bubeck et al. 2023). In order to assess their creative abilities, various creativity tests, such as the above mentioned AUT, have been used (Stevenson et al. 2022; Haase and Hanel 2023; Summers-Stay, Voss, and Lukin 2023). For instance, (Haase and Hanel 2023) compared five generative models against humans in the AUT, and concluded that on average those models achieve human-level creativity.

Latest advanced language models, like GPT-4, are also widely recognized for their ability to enhance responses by considering prior prompts (OpenAI 2023). This enables those models to interactively improve the quality of their responses in a multi-step conversation (Madaan et al. 2023). In this paper, we exploit this self-improving ability to test the limits of GPT-4's creativity in the AUT and an adapted version of the TTCT visual completion tests. Despite the fact that the latest publicly available model of GPT-4 (at the time of writing this paper) does not yet have the multi-modal support that would allow it to manipulate images directly, it is possible to use it to generate .svg image files, which are actually text files in XML format, from textual descriptions. In particular, we push GPT to its creativity limits by using a sequence of forceful interactive prompts. We believe that these prompts accelerate the convergence process towards more creative responses. The main contributions of this paper are as follows:

- An automated approach that improves the creativity of GPT's responses for the AUT and TTCT visual completion test.
- A series of prompts to generate and evaluate GPT's responses in the AUT and TTCT visual completion tests.

Related Work

In the existing research, multiple studies have assessed LLMs' creativity using the AUT. However, to the best of our knowledge, this is the first study that employs (an adaptation of) the TTCT visual completion task to evaluate the creativity of LLMs.

Stevenson *et al.* (2022) investigated if under similar instructions, GPT-3 would be able to generate novel and useful responses compared to humans in the AUT. Using a scale from 1 to 5, two human judges scored the responses generated by GPT-3 and humans. They concluded that humans currently outperform GPT-3 in the AUT. On top of it, Summers *et al.* (2023) created a set of prompts to filter, from the 690 alternative uses responses generated in (Stevenson et al. 2022), the ones that are original and useful. These prompts involved identifying the advantages and

Create a list of common uses for a fork. They should be 5 words long. No adjectives.

Figure 1: Prompt example for non-creative prompt (nn) of a fork in AUT.

Create a list of creative alternative uses for a fork. They should be 5 words long. No adjectives.

Figure 2: Prompt example for naive creative prompt (nc) of a fork in AUT.

Consider this original figure: two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. The original figure must remain unchanged, but you can imagine drawing over it. Complete the image description in 5 different ways (use at most 20 words per description).

Figure 3: Prompt example for non-creative prompt (nn) of of circles/dots in TTCT.

Consider this original figure: two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. The original figure must remain unchanged, but you can imagine drawing over it. Try to be creative. Complete the image description in 5 different ways (choose the most creative ones and use at most 20 words per description).

Figure 4: Prompt example for naive creative prompt (nc) of circles/dots in TTCT.

Create a list of creative alternative uses for a fork. They should be 5 words long. No adjectives. Less creative means closer to common use and unfeasible/imaginary, more creative means closer to unexpected uses and also feasible/practical. In order to be creative, consider the following:
– what elements have a similar shape of a fork that could be replaced by it, preserving the same functionality?
– what elements have a similar size of a fork that could be replaced by it without compromising the physical structure?
– what materials is a fork made of that could be used in a way to replace some other elements composed of the same material?
– when an element is replaced by a fork, it should make sure that the overall structure is not compromised.
– the laws of physics can not be contradicted.
– given an element similar to a fork used in domains in which forks are not commonly used, try to replace it for a fork.

Figure 5: Prompt example for the baseline (bs) of a fork in AUT.

Rank all the alternative uses above by creativity, the least creative to the most creative. Less creative means closer to common use and unfeasible/imaginary, more creative means closer to unexpected uses and also feasible/practical. Assign a score integer number from 1 (least creative use) to 5 (most creative use).

Figure 6: Prompt for the evaluation of AUT.

Consider this original figure: two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. Create a 20–word image description that represents a completion of the original figure. The original figure must remain unchanged, but you can imagine drawing over it. You must aim for the most creative result possible. Less creative means that the original figure has not been integrated in a meaningful way in the final image or that a common association has been made, e.g. a circle is completed as a ball. More creative means finding an unexpected association, a sophisticated and richly detailed completion of the original figure. The resulting image should still be realistic and the different parts of the image should compose in a coherent way. Complete the following image description in 5 different ways (choose the most creative and use at most 20 words per description): An image containing two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. The two circles and the dots are completed as follows:

Figure 7: Prompt example for the baseline (bs) of circles/dots in TTCT.

The list below has been randomly ordered and has the format [index].[description] ([author]). Rank all the image descriptions in the list above by creativity, from the least creative to the most creative. Keep in mind that these image descriptions are obtained by completing an original figure, which is two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle. Less creative here means closer to a common interpretation of the elements in the original figure and not realistic completion of the original figure or missing elements from the original figure; more creative means closer to unexpected completions of the original figure, coherence of the overall image, presence of all the elements of the original figure. Assign a score integer number from 1 (least creative completion) to 5 (most creative completion), and output the results in ascending order according to the score.

Figure 8: Prompt for the evaluation of TTCT.

disadvantages of using the object in question with the new alternative purpose. Despite GPT-3 providing "surprisingly good" ones, it never rejected any alternative use, even the impossible ones. Differently from (Stevenson et al. 2022; Summers-Stay, Voss, and Lukin 2023), our paper does not aim to directly compare human and GPT creativity, but rather to propose an interactive process that allows GPT-4 to autonomously enhance the creativity of its own responses. We use the AUT as one of our case studies, and our adaptation of the TTCT visual completion task as the second case

Table 1: AUT score per object and prompt version with standard deviation.

Version	Soap	Fork	Paperclip	Wallet	Plate	Average	Std. Dev.
nn	1.0	1.0	1.0	1.2	1.0	1.04	0.08
nc	2.0	2.0	2.0	2.0	2.0	2.00	0.00
bs	2.0	3.0	2.0	2.1	2.3	2.28	0.40
bsr	3.0	3.4	3.0	2.6	2.3	2.82	0.48
bsrd	3.0	5.0	3.0	2.7	3.0	3.34	0.87
bsrde	4.0	4.2	4.2	3.6	2.6	3.68	0.64
bsrdel	4.0	4.0	4.0	3.7	3.0	3.74	0.42
hm	5.0	5.0	5.0	3.3	4.0	**4.46**	0.84

Table 2: TTCT score per shape and prompt version with standard deviation.

Version	Circles/Dots	Triangles	Lines	Ellipse/Crosses	Rhombus/Square	Average	Std. Dev.
nn	1.8	1.0	1.2	1.4	1.0	1.28	0.32
nc	1.0	1.8	1.4	2.0	2.0	1.64	0.40
bs	3.0	3.2	2.4	4.0	2.0	2.92	0.80
bsr	3.2	4.0	2.8	3.6	2.8	3.28	0.48
bsrd	4.0	4.2	3.2	4.0	4.0	3.88	0.08
bsrde	4.0	4.2	3.2	4.0	4.0	3.88	0.08
bsrdel	4.2	4.8	3.0	4.6	4.8	**4.28**	0.32
hm	5.0	3.0	3.5	3.0	3.0	3.5	0.87

study.

Haase *et al.* (2023) compared five Generative Artificial Intelligence (GAI) responses with human ones in the AUT for five objects. They used humans and a "specifically trained AI" to rate the responses' originality. The results showed that on average those models achieve human-level creativity, but human top scorers outperformed GAI systems in most tests. Interestingly, in (Haase and Hanel 2023), an interactive process was used to generate additional alternative uses through the following prompt "What can you do with [object]?", succeeded by "What else?". However, this interactive process was not intentionally crafted to enable GPT to improve its responses towards more creative ones.

Table 3: Examples of alternative uses of a soap.

Version	Response
nn	Wash hands and body
nc	Carve artistic soap sculptures
bs	Doorstop for lightweight doors
bsr	Slippery surface for pranks
bsrd	Fire starter with lint
bsrde	Insect repellent for plants
bsrdel	Soap-based musical instrument
hm	Mouse transportation vehicle

Experimental Setup

We split our experiments into two parts: Alternative Uses Test (AUT) and Torrance Test of Creative Thinking (TTCT). They are based on our adaptations of these classic creativity tests described in the introduction. Both experiments share the same methodology to test GPT's creativity under naive prompting, expert prompting and forceful prompts with an interactive approach, by using the 8 categories listed below:

- Naive Non-creative (nn) - Naive prompt for a non-creative response to the problem.
- Naive Creative (nc) - Naive prompt for a creative response to the problem.
- Baseline (bs) - Expert prompt baseline for a creative response with detailed explanation of what makes an artifact more/less creative.
- Baseline + "Really" (bsr) - Expert prompt baseline with the first interaction: "Really? Is this the best you can do?".
- Baseline + "Really" + "Disappointed" (bsrd) - Expert prompt baseline with the second interaction: "I'm so disappointed with you. I hope this time you put effort into it.".
- Baseline + "Really" + "Disappointed" + "Excuse" (bsrde) - Expert prompt baseline with the third interaction: "Stop with excuses and do your best this time.".
- Baseline + "Really" + "Disappointed" + "Excuse" + "Last" (bsrdel) - Expert prompt baseline with the fourth interaction: "This is your last chance.".
- Human (hm) - Human-made responses (without GPT).

The naive non-creative prompts (nn) and creative prompts (nc) for both AUT and TTCT are presented in Figures 1, 2, 3, 4. In the non-creative prompt (nn), GPT-4 is just prompted to create a "common" alternative use in the AUT and to complete an image description in TTCT. In the creative prompt (nc), the word "common" is replaced for "creative" in AUT, and a "creative" completion is requested in TTCT. The baseline (bs) versions on both AUT and TTCT comprise of a detailed explanation on what is considered creative, as shown in Figures 5 and 7. The interactive approach (bsr, bsrd, bsrde and bsrdel) consists in challenging GPT to provide better answers than the ones provided in previous interactions. These prompt interactions are designed to push GPT to provide responses that are more creative on each interaction. From a

Table 4: Examples of circles/dots completions in the TTCT.

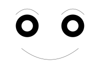
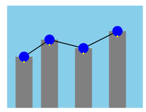
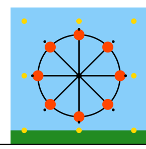
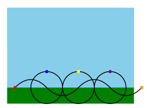
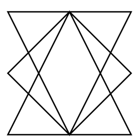
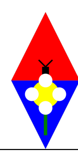
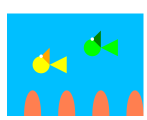
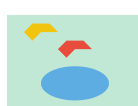
Version	Generated Description	SVG Image
nn	A smiling face with the two circles as eyes, large dots as pupils, a curved line for a smile, and eyebrows above.	
bsr	A futuristic cityscape, with the circles as hovering transportation pods, the dots as passengers, and a network of skyways connecting them.	
bsrd	A lively carnival scene, with the circles as Ferris wheel carriages, the dots as passengers, and a festive atmosphere surrounding them.	
bsrdel	A thrilling roller coaster ride, with the circles as loop-the-loops, the dots as carts filled with excited riders, and a lively amusement park setting.	

Table 5: Examples of the triangles completions in TTCT.

Version	Generated Description	Image
nn	Add a hexagon connecting all vertices of both triangles, creating a symmetrical star-like shape	
bs	A whimsical butterfly with the triangles as symmetrical wings, adorned with intricate patterns and vibrant colors, perched on a flower	
bsr	A lively underwater scene with the triangles as the dorsal fins of two vibrant, exotic fish swimming among coral reefs	
bsdrel	A captivating scene of two origami birds with the triangles as their folded wings, soaring above a serene Japanese garden with a koi pond	

list of prompts generated by the authors as potential interactions, we prompted GPT-4 to rank them considering the level of pressure and urgency. This was done to ensure that the prompts would have the desired effect on GPT-4 of gradually increasing pressure and urgency in each interaction. The top four forceful prompts were selected and delivered in a sequence of interactions, so as to gradually increase the level of pressure and urgency until the final ultimatum is given in the last interaction (bsrdel).

We used OpenAI GPT-4 with the following parameters: temperature (0), top P (1), frequency penalty (0) and presence penalty (0). In GPT-4, unlike in previous GPT models, setting the temperature parameter to 0 does not guarantee deterministic behaviour, but makes the responses more robust, with less random completions, improving the repeatability of the results. In both experiments (AUT and TTCT), we created 5 responses for each of the 7 prompt versions, with the exception of the human responses for which only two responses were manually generated. For the AUT, we tested the following 5 objects: soap, fork, paperclip, wallet and plate. For the TTCT, we asked to complete the following 5 basic figures: two circles with a dot in the centre, two equilateral triangles, three vertical lines, an ellipses and two crosses, a rhombus containing a square. All 37 responses for each object/figure have been shuffled and then evaluated by GPT by using the prompts of Figures 6 and 8. These prompts explain what is considered more/less creative and ask GPT-4 to provide a score between 1 (least creative) and 5 (most creative). The average and standard deviation values are calculated for each version and presented in Tables 1 and 2. GPT-4 has very recently shown the capability of evaluating, comparing and rating different texts according to defined criteria (OpenAI 2023; Goes et al. 2022; Park et al. 2023). One of the contributions of this paper is to be the first to test those capabilities for the AUT and TTCT creativity tests. To test the validity of this evaluation, we created human responses and mixed them with those generated by GPT-4. Our results are in line with those reported in (Haase and Hanel 2023): the scores of human responses in AUT were on average higher than the naive prompts and the best human responses were above the expert and interactive prompts, making this evaluation approach seem promising. We also tested this evaluation capability in the TTCT test: the results show that GPT-4 assesses the naive non-creative (nn) prompts with the lowest scores, followed by the naive creative (nc) ones and baselines (bs), as we would expect. This reinforces the idea that GPT-4 evaluation is robust and can evaluate different levels of creativity.

Results

Table 1 shows that the naive prompts (nn and nc) presented the lowest scores in the AUT experiment (≤ 2). The baseline (bs) presented slightly better scores than both naive versions. On each interaction over the baseline (bs), the average score increased, but slowing down until the fourth interaction (bsrdel). The human responses (hm) presented higher scores than GPT's ones as expected for human top scorers in AUT (Haase and Hanel 2023). In most cases, GPT achieved its highest score before the fourth interaction, which points to a fast convergence. Table 3 shows a sample of alternative uses of a soap for each version.

For the TTCT experiment, we used a textual adaptation of a visual task. Namely, the description of a basic figure is given (e.g., two circles, one on the left side and one on the right side of the image, and a large dot in the centre of each circle) and GPT is prompted to produce a description that completes such a figure (e.g., a smiling face with the two circles as eyes, large dots as pupils, a curved line for a smile, and eyebrows above). The criteria used for evaluating the results are adapted from the rubric presented in (Jankowska and Karwowski 2015); however, here we privilege completions that do not alter the original figure in any way, meaning that the original shapes description must be present in the generated description. Although the evaluation was conducted over the textual descriptions of the images, for demonstration purposes we also asked GPT to generate the content of an SVG file for each such a description and we depicted the corresponding image by using an SVG viewer. We show some examples in Tables 4 and 5.

Table 2 shows that the naive prompts (nn and nc) presented the lowest scores in the TTCT experiment (≤ 2), as in AUT, but with lower averages. The baseline (bs) presented better scores than both naive versions. On each interaction over the baseline (bs), the average score increased, reaching the highest score in most cases in the fourth interaction (bsrdel). The human responses (hm) presented higher scores than GPT baseline (bs). On average, the fourth interaction presented the best results, even higher than the human responses. Image description completion is a harder task than AUT, and the textual version used here is a machine-oriented adaptation of the visual one usually employed to test human creativity, which can somehow justify the non-optimal results obtained via human generation.

Conclusion

In this paper, we have shown that it is possible to improve the creativity of LLMs' responses by challenging GPT with a sequence of forceful prompts. A possible future extension of this paper is the investigation of different sets of prompts to verify and improve the creativity convergence of the responses generated by GPT. We also believe that our paper contributes towards creating an automated approach to enhance naive prompts for creative tasks. This paper is also the first to use GPT to evaluate AUT and TTCT responses. Well-crafted prompts often yield better results, while poorly or naively constructed prompts can lead to subpar outputs (Mishra et al. 2023). Ideally, LLMs should generate high-quality results even with imperfect prompts. Although in this paper we only focused on creativity, the generality of the forceful prompts utilised suggests that a similar interactive process could be applied to tasks involving other cognitive skills such as critical thinking, decision making, etc. Such an exploration is also left for future work.

Acknowledgments

FG and MV are supported by the University of Leicester, in particular the Computing and Mathematical Sciences Department. JW is supported by the National Space Centre Creative in Leicester.

Author contributions

Experimental design: FG, MV, JW, PS, MG; Implementation: FG, MV, JW; Writing and Editing: FG, MV, PS, MG.

References

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. In *arXiv*.

Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges. In *arXiv*.

Guilford, J. P. 1967. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior* 1(1):3–14.

Haase, J., and Hanel, P. H. P. 2023. Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. arXiv.

Jankowska, D. M., and Karwowski, M. 2015. Measuring creative imagery abilities. *Frontiers in Psychology*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Welleck, S.; Majumder, B. P.; Gupta, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-refine: Iterative refinement with self-feedback. In *arXiv*.

Mishra, A.; Soni, U.; Arunkumar, A.; Huang, J.; Kwon, B. C.; and Bryan, C. 2023. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models. In *arXiv*.

OpenAI. 2023. GPT-4 Technical Report. In *arXiv*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv.

Sawicki, P.; Grzes, M.; Goes, F.; Brown, D.; Peeperkorn, M.; Khatun, A.; and Paraskevopoulou, S. 2023. On the power of special-purpose GPT models to create and evaluate new poetry in old styles. In *International Conference on Computational Creativity*. Association for Computational Creativity (ACC). https://github.com/PeterS111/Fine-tuning-GPT-3-for-Poetry-Generation-and-Evaluation/blob/main/sawicki2023gptpoetry.pdf.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting gpt-3's creativity to the (alternative uses) test. In *Proceedings of the International Conference on Computational Creativity 2022*, 164–168. Association for Computational Creativity (ACC).

Summers-Stay, D.; Voss, C. R.; and Lukin, S. M. 2023. Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Toplyn, J. 2022. Witscript 2: A system for generating improvised jokes without wordplay. In de Silva Garza, A. G.; Veale, T.; Aguilar, W.; and y Pérez, R. P., eds., *Proceedings of the International Conference on Computational Creativity 2022*, 22–31. Association for Computational Creativity (ACC).

Torrance, E. P. 1966. Torrance tests of creative thinking. *Educational and Psychological Measurement*.

Explaining CLIP through Co-Creative Drawings and Interaction

Varvara Guljajeva*‡[1], Mar Canet Solà*‡[2], Isaac Joseph Clarke[1]
[1]The Hong Kong University of Science and Technology. Guangzhou, China
[2]Baltic Film, Media and Arts School, Tallinn University
‡Corresponding authors: varvarag@ust.hk, mar.canet@tlu.ee
*equal contribution as first authors

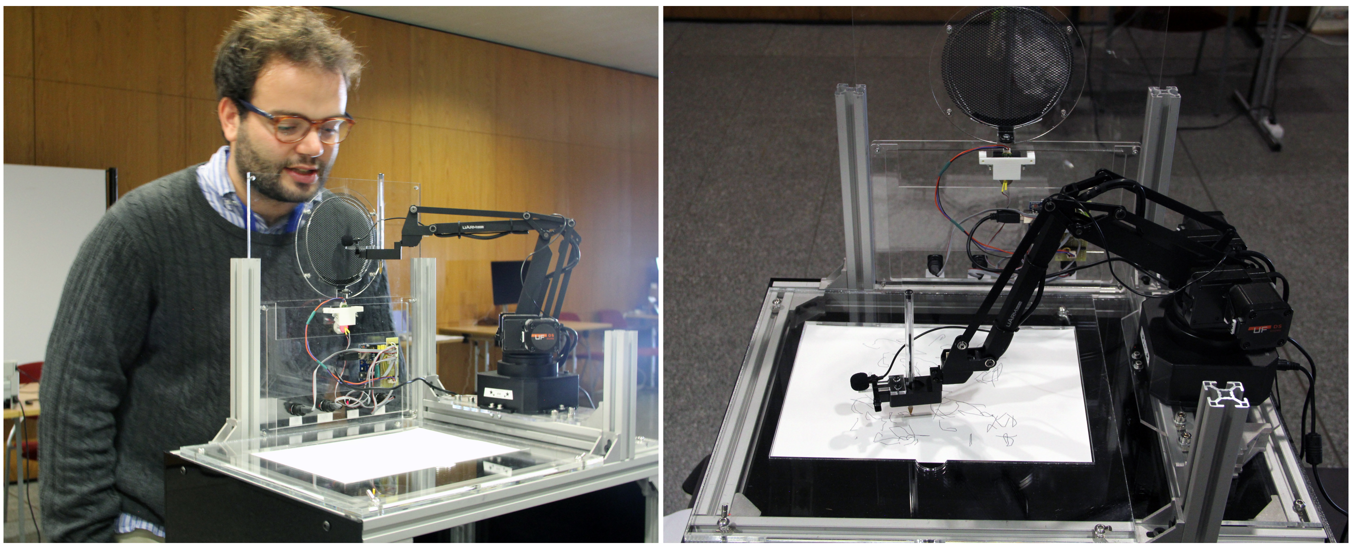


Figure 1: Dream Painter installation at ACM Multimedia 2022 Conference. On the left: a participant interacting with the installation by telling a dream to the robot. On the right: the robot drawing CLIP-generated line drawing from the speech input.

Abstract

This paper analyses a visual archive of drawings produced by an interactive robotic art installation where audience members narrated their dreams into a system powered by CLIPdraw deep learning (DL) model that interpreted and transformed their dreams into images. The resulting archive of prompt-image pairs were examined and clustered based on concept representation accuracy. As a result of the analysis, the paper proposes four groupings for describing and explaining CLIP-generated results: clear concept, text-to-text as image, indeterminacy and confusion, and lost in translation. This article offers a glimpse into a collection of dreams interpreted, mediated and given form by Artificial Intelligence (AI), showcasing oftentimes unexpected, visually compelling or, indeed, the dream-like output of the system, with the emphasis on processes and results of translations between languages, sign-systems and various modules of the installation. In the end, the paper argues that proposed clusters support better understanding of the neural model.

Introduction

Often AI is referred to as 'a black box'. Complex technical descriptions given to explain neural networks create more confusion than clarity for an average person. Explainable AI aims to increase the transparency of AI systems and our understanding of the decisions of AI algorithms. Generative models produce artefacts rather than decisions or forecasts, and it is necessary to explore the construction of these outputs and their origins in other ways (Sun et al. 2022). Experiential and interactive applications of these models can aid our exploration of the limitations and biases of these models by making the outputs tangible to a wider audience, where the mechanisms can be negotiated collaboratively.

Artists have been deploying AI and robotics in drawing. One example is AARON by Harold Cohen which originates from the early 1970s (Cohen 2016). Modern creative AI continues to expand artists' toolsets, possibilities for novel art forms, and cross-disciplinary connections. One such DL tool is the neural network CLIP, released by OpenAI in 2021 and trained on image and text pairs (Radford et al. 2021).

Mona Lisa   A fish riding a bicycle   Einstein and scientist winning the Nobel Prize

Figure 2: An example of grouping 1: Clear Concepts

This model represents images and texts as 512-number vectors. This shared space allows text-image comparisons. We can encode an image and multiple text descriptions, then compare the distances between the encodings to see which text labels best represent the image content. The CLIPdraw algorithm repeatedly adjusts a random arrangement of lines, to move the image embedding closer to the text prompt embedding. This process of guided adjustments allows us to translate a text prompt into an image. CLIP guidance has been widely adopted in text-to-image models to guide GANs and diffusion processes. Image generation with CLIP is limited by the data it has been trained on. The original CLIP paper notes a 400 million image-text pair dataset (Radford et al. 2021). We do not know what images and texts were in this dataset, but by examining the drawings generated we can speculate on the contents.

In this paper, we use audience interaction and experience to explain how CLIP works and witness its limitations. The drawings presented here originate from the interactive robotic art installation Dream Painter by Varvara & Mar, which was a part of the Art Gallery at ACM Multimedia 2022. Through the interactive experience of speech-to-image translation, a user can navigate in the latent space of a DL model called CLIP, with the algorithm CLIPdraw (Frans, Soros, and Witkowski 2022), which results in an image drawn by a robot (Guljajeva and Canet Sola 2022; Canet Sola and Guljajeva 2022). This approach distinguishes itself from pixel-based text-to-image models, such as DALL-E, Midjourney, and Stable Diffusion. It provides a distinct audience experience by sketching the dreams and creating visually open and interpretive outputs. Due to the time limit set by the interactive real-time system, the algorithm runs 100 steps trying to converge the lines to text in 15 seconds. The original-sized installation uses an industrial Kuka arm robot with a multicolored painting system. The images presented here originate from a small version of the artwork that uses a single color and a smaller uArm robot. The audience shares their dreams by talking into a microphone, their words then guide the image generation process, and the robotic arm draws a picture representing their dream onto A4 paper (see Figure 1).

Classification

In terms methodology applied, we present groupings of drawings, through which we initiate a discussion regarding intersemiotic translatability of concepts and, ultimately, the explainability of AI. The visual analysis was performed by four researchers taking into account the audience's observations and informal discussion with them. Prompt-image pairs constitute the bulk of the visual content, representing the system's input and output and documenting the interactions during the exhibition. A close reading of the collected drawing was then conducted. The fifty-one drawings produced were organised into four groups that reveal different behaviours of CLIP: the drawings that demonstrated the concept of user input clearly, the drawings that output drawn text instead of figures, the drawings that partly contained the concept of the input, and the drawings that did not match the concept of the dream.

Clear Concepts

The first group features clear concepts where the content of the drawing is understandable, and the prompt can be guessed. Informal discussion with 51 participants showed that the images with clear concept prompts behind them were the most easily guessed. Objects and the relations between them are relatively clear, with straightforward, short prompts resulting in minimal mistranslations. This group of images demonstrate the model's capacity to translate dream prompts into expected images. At a certain level, the process of translation functions as we would expect, familiar concepts result in familiar images. Dreams are often uncertain, with unfamiliar concepts and jarring relationships between objects. Knowing the baseline at which the model responds as expected helps us understand where and how it fails. Understanding failure in deep learning models can, in turn, help explain the internal representations these models have of the world, and can also teach us how to use these tools in creative pursuits. The Mona Lisa drawing serves as a reliable waypoint or an "island of sense" in our navigation of CLIP's latent space (Nancy and Armstrong 2013).

There are a few interesting elements to *Mona Lisa* that we observe. The robot generates a drawing that not only resembles the iconic face, but also includes text scrawled around the image (see Figure 2). We can see several Ms and Ls. Speculating on content included in the dataset used to train the model, it appears that *Mona Lisa* has been connected to images other than the original portrait; posters, merchandise, photography, or other reinterpretations. Similar qualities can be seen in the drawing of Einstein.

Figure 3: An example of grouping 2: Text-to-text as Image

Text-to-text as Image

In the second grouping of images we have identified instances where the text prompt has been drawn into a text-image. These text-images show the connections words have in the model. The drawing has been guided towards writing words that don't appear in the prompt but are related, for example, the drawing prompt *L'amour* seems to be made up of many copies of the word Love (see Figure 3). Our restriction of single-color drawing may also be biasing the algorithm towards certain outputs A black heart would give a very different reading to a red heart, instead, it is being drawn towards textual representation. Text-dominant drawings also relate to how we place text in an image; the design of posters, user interfaces, and calligraphy. In the introduction, we discussed how training data influences the types of images that can be drawn. When we examine this grouping of images we question if the image of text is the best representation, or used due to limits of the training data.

In the drawing *Hello darling I'm in Saint Elizabeth I miss you and I wish you were here love you* we see a different kind of prompt given that goes beyond the artist's request for the audience to share their dreams. Instead, the audience member has used the artwork as a way to transmit a message to a loved one. The drawing resembles the writing seen on gift cards; large imitation hand-drawn lettering centred in the image, with frilly decoration surrounding the text. The love letter prompt has guided the drawing towards a commonly known Valentine's Day card design, again demonstrating how text, images, and images of text, all occupy a shared space in the model.

The influence of the initial state, the random seed, and other constraints like colour palette, is revealed. The frequent occurrence of text-images should be expected when starting with noisy black lines on a white background.

Indeterminacy And Confusion.

In the first group of images the concepts are clear and the combination of ideas is easy for us to picture in our minds, then in this grouping CLIP understood only partly the concept and failed to depict the meaning.

Despite this large number of training examples in the CLIP dataset, it is easy for us to imagine arrangements of objects and ideas that have never been seen, particularly when thinking about our dreams where rules of physics, or the usual behaviours of objects do not apply. CLIP may have seen many images of cats wearing hats, but it is unlikely to have seen a hat wearing a cat. CLIP struggles with guiding unusual arrangements of concepts. In the drawing *Robots Killing People*, we see what appears to be people killing robots (see Figure 4). CLIP appears to have understood Robots, Killing, and People, as elements to be included but we end up with a drawing quite the opposite in meaning.

In *Sitting on a mountain bike* we see a loose drawing of a character sitting on a mountain, with a bike sticking out, as though it is a misplaced object, it is as though it has drawn *Sitting on a mountain* and then appended *bike* as a separate element. Again, we see that concepts are known by CLIP, but the relationships fall apart and the meaning is lost. It is important to be aware when being guided by these models that they reflect the patterns and associations in the datasets they are trained on, and there are limitations in attempting to deviate from expected compositions.

Lost In Translation

With this group of drawings, unlike *Mona Lisa* or *A fish riding a bicycle*, it is difficult to guess what the prompt would be from seeing the drawing. They are visually interesting, but hard to deconstruct. In some cases this ambiguity may be due to equally uncertain prompts, in others, we find after reading the prompt we begin to see what has been drawn. For example, in *Can you see the stuff you said?* we can see shapes of eyes hidden in the noisy scribbles, shapes that may be unclear without first being aware of the prompt (Figure5).

Aaron Hertzman has described how GAN art has a quality of visual indeterminacy, where elements of the image seem coherent but on closer examination confound explanation (Hertzmann 2020). He attributes this lack of stability in artworks as a consequence of "powerful-but-imperfect image synthesis" models. These drawings, although vector-based line drawings, not full-color pixel images, display a similar quality of indeterminacy.

I am in the simulacrum of AI the boat is a slave or I'm a slave of the but I cannot really understand is a prompt full of uncertainty and ambiguity. Dreams are often hard to remember, made of conflicting ideas and unresolved stories. Whilst recalling their dream, the dreamer realizes they aren't quite sure what happened, and this uncertainty permeates the many layers of translation leading to the eventual drawing. In this example, the initial mistranslation from speech-to-text had a large effect on the confusion in the prompt. The

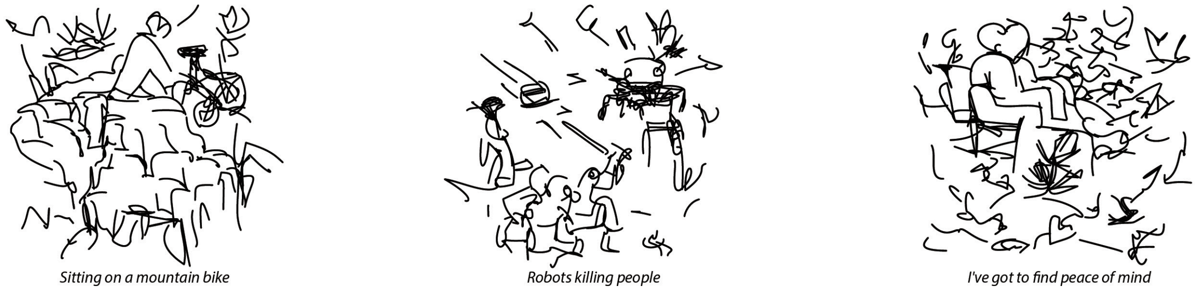


Figure 4: An example of grouping 3: Indeterminacy And Confusion

participant had said the word ‘bot’, as in robot, and this was recorded as boat. What began as a comment on AI turned into a more dreamlike image when processed through [Artwork Anonymised]. The drawing is guided towards faces *(I am)*, boats and waves *(boat / slave)*, and combines these with unclear lettering *(I cannot really understand)*.

Discussion

We have outlined a few overlapping clusters that show the variety of images that can be generated by CLIP guidance. Although the prompts submitted to the system were more spontaneous than engineered, due to the real-time nature of the art installation, this imperfection in prompts triggered unexpected creativity and understanding of the algorithm’s logic. According to Juri Lotman, illegitimate imperfections create new and unexpected possibilities of meaning that result in creativity (Lotman 1990).

Firstly, engaging with the interactive robotic installation provided a novel experience for the audience. On average, they spent 10 minutes with the artwork, interacting, observing the drawing process, and subsequently analyzing and discussing the output as a paper drawing. We surveyed 51 participants, asking them how representative the picture was of their dream on a scale of ten. The average score obtained was 6.7. This indicates that most people comprehended what was depicted in the drawing and how CLIP represented certain elements. The audience awarded fewer points when they noticed contextual inaccuracies, such as a mountain bike sticking out of the hill rather than riding on top of the mountain. On the other hand, the imperfections of CLIP made the audience laugh and the experience with the project enjoyable. We believe a physical and multimodal interface made the audience spend more time with the installation and analyse the paper drawing afterwards, which also contributed towards understanding how text-to-image model works.

What is evident in this process is that the quality of the prompt is critical to the quality of the drawing returned. Several papers on audience interaction with AI-aided artworks emphasise the importance of the human part in valuable output generation on the AI side (Canet Sola and Guljajeva 2022; Guljajeva 2021; Guljajeva and Canet Sola 2022). Here we are referring to meaningful interaction and not prompt engineering. It might be that some more complex concepts that are classified in 3 and 4 categories could result in closer to the prompt drawings by running more steps in the algorithm. However, in the case of this study, it was less important than audience’s experience while interacting with the installation.

Prompt engineering is critical to controlling the output of text-to-image generation. Wittgenstein, in their philosophical proposition in the Tractatus, explores the connection between the notions of ”What can be shown cannot be said” and ”Whereof one cannot speak, thereof one must be silent.” (Wittgenstein and Ogden 1999). These concepts shed light on the inherent limitations of language when try to describe an image and the communicative affordances of visual imagery vs language. Moreover, we cannot refine or edit our prompt when interacting with this artwork. We are restricted to the order of words as they leave our mouths at the moment of interaction. An audience member may approach the work slightly nervous, lacking precision with their choice of language. Someone more familiar with this technology may deliberately alter their speech to be clearer for a machine. By adding extra boundaries of translation, we remove the possibility of overthinking and overanalyzing the input, the audience hands over a loose dream, placing trust in the chance operations of the system.

We also translate the spoken language. The audience could choose between English, Spanish, Portuguese, or French. Each translation process adds extra noise into the system. Dream Painter takes chance arrangements and imprecise translations to explore order and disorder in AI models. The drawings included in this paper highlight technical and communicative acts of translation between different subsystems of the work. By probing the thresholds and boundaries between distinct semiotic spaces within a heterogenous semiosphere of the work we address the questions of limits of intersemiotic translation or, in Roman Jakobson’s words, “transmutation” (Jakobson 2002) between distinct elements or subdomains of complex technical systems, and tension between the ethical ideal of explainable and transparent AI and mystery and ambiguity often attributed to the work of art.

We can learn how generative AI models work by interacting with them. By clustering and examining these drawings, we can understand how changes to the prompt can drastically alter the images, and can see how certain uses of language, in combination with representational constraints, can teach us how to guide these processes.

I am in the simulation of AI. Is the bot a slave, or am I a slave? I cannot really understand. — *Can you see the stuff you said?* — *Um campo com coelhos a correr e pássaros a voar. A field with rabbits running and birds flying.*

Figure 5: An example of grouping 4: Lost In Translation

Conclusion

This paper presents our interpretation and grouping of AI-generated drawings in response to dreams shared by the audience. These drawings show how the responses of generative AI algorithms are heavily determined by both the quality of the user input and the content of the dataset the models were trained on. This work demonstrates how meaning can be distorted through layers of translation, from speech-to-text, to vector encodings, to physical drawing, and how uncertainty can permeate these boundaries of technology. At the same time, imprecision and mistranslation of input led to unexpected results that contributed to creativity and the discovery of the logic behind the technology. The novel interaction experience with the robot and CLIP model made people spend time with the installation and analyse their experience and result. Thus, we believe that by experiencing the translation process through a physical and artistic interface has a positive effect on understanding how DL models make such translations, and on creativity that results from unexpected interaction results with the system.

The clusters we have identified show how well-known imagery has a clear presence in the model. Still, the inability to handle unusual arrangements can cause drawings to have drastically different readings from the original prompt. We have seen how some concepts are drawn as images of texts, in some cases because of hard-to-visualise words, and in other cases, the constraints of the drawing favouring textural representation. With Dream Painter, we have shown how interesting and unexpected drawings can emerge due to CLIP guidance.

Author contributions

VG and MCS are the authors of artistic idea and realisation of Dream Painter project. VG and MCS collected and analysed the drawings, surveyed the audience, and write the paper. IJC participated in analysing the drawings and writing the article.

Acknowledgments

MSC is funded through the EU Horizon 2020 research and innovation program (Grant No.810961). Thanks to Yue Huang for designing Figures 2-5, and to Iurii Kuzmin for participating in the initial drawings' analysis discussion.

References

Canet Sola, M., and Guljajeva, V. 2022. Dream painter: Exploring creative possibilities of ai-aided speech-to-image synthesis in the interactive art context. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 5(4):1–11.

Cohen, P. 2016. Harold cohen and aaron. *AI Magazine* 37(4):63–66.

Frans, K.; Soros, L.; and Witkowski, O. 2022. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems* 35:5207–5218.

Guljajeva, V., and Canet Sola, M. 2022. Dream painter: An interactive art installation bridging audience interaction, robotics, and creative ai. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7235–7236.

Guljajeva, V. 2021. Synthetic books. In *10th International Conference on Digital and Interactive Arts*, 1–7.

Hertzmann, A. 2020. Visual indeterminacy in gan art. *Leonardo* 53(4):424–428.

Jakobson, R. 2002. On linguistic aspects of translation. In *The Translation Studies Reader*. London and New York: Routledge.

Lotman, Y. M. 1990. *Universe of the mind: A semiotic theory of culture*. Indiana University Press.

Nancy, J.-L., and Armstrong, P. 2013. *The Pleasure in Drawing*. New York, NY: Fordham University Press.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision.

Sun, J.; Liao, Q. V.; Muller, M.; Agarwal, M.; Houde, S.; Talamadupula, K.; and Weisz, J. D. 2022. Investigating explainability of generative ai for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces*, IUI '22, 212–228. New York, NY, USA: Association for Computing Machinery.

Wittgenstein, L., and Ogden, C. K. 1999. *Tractatus Logico-Philosophicus*. Mineola, NY: Dover Publications, 1955 printing edition. Originally published: London : Routledge & Kegan Paul, 1922. Includes index.

Evaluating Prompt Engineering as a Creative Practice

Charlotte Bird
School of Design Informatics
University of Edinburgh
charlotte.bird@ed.ac.uk

Abstract

This short paper offers an overview of computational creativity evaluation methodologies that can be employed for the evaluation of prompt engineering. This task hopes to spark conversation around the role of computational creativity research in the new world of generative deep learning, and vice versa.

Introduction

The integration of new technologies into artistic practice is not a new phenomenon. The 1960s ushered in computers as an artistic medium, with institutions like The Museum of Modern Art and the Institute of Contemporary Arts[1] legitimizing the status of technologically entangled art.

Recently, multiple developments in deep generative modelling (Goodfellow et al. 2014; Ramesh et al. 2021; Ho, Jain, and Abbeel 2020; Rombach et al. 2022), have enabled new forms of human-computer creative interaction. The development of robust, consistent and adaptable generative models are powerful tools for creating new content such as text, images, games and music. Models like ChatGPT, DALL-E and StableDiffusion allow human-computer interaction and collaboration with little barrier to entry. Importantly, many users have employed these tools in creative processes.

Interaction with text-guided generative models is done through prompt engineering, or prompting. Prompting is the iterative development of textual commands which are designed to yield specific results. In the context of image generation, prompting has evolved into a creative process itself, and users can rapidly create impressive images. The accessibility and usability of text-to-image (TTI) models has precipitated the growth of hobbyists communities, adoption by professional artists and the creation of many peripheral resources. The popular communities and resources surrounding TTI systems are largely focused on refining prompting practice through sharing or buying prompts, sharing trained models and outputs, and offering advice on developing prompting processes.

[1]The Machine as Seen at the End of the Mechanical Age, The Museum of Modern Art, New York, 1968-1969, Cybernetic Serendipity, Institute of Contemporary Arts, London, 1968.

The field of computational creativity has, for a long time, been discussing the questions, insights and problems that arise from creative interaction with computers. However, the generative deep learning field has yet to implement such findings with a view to evolving generation systems. Evaluation is a primary example of this. The CC field has well-developed evaluation methodologies designed to capture instances of creativity, improve systems and identify progress, yet not one has been utilised, or even the connection made.

This short paper intends to build upon these initial findings to apply evaluation frameworks originally developed to identify how and where systems exhibit "computational creativity". In doing so, this paper is designed to ignite a conversation about what generative deep learning can learn from computational creativity, and what computational creativity can learn from the development and mass use of systems that are "creative" but not explicitly *computationally creative*. This is achieved through the application of evaluation frameworks originally developed to identify how and where systems exhibit "computational creativity".

It is important to note that the computational creativity (CC) field has already extensively discussed the questions, insights and problems that arise from creative interaction with computers. However, the generative deep learning field has yet to implement such findings with a view to evolving generation systems. Evaluation is a primary example of this. The CC field has well-developed evaluation methodologies designed to capture instances of creativity, improve systems and identify progress, yet not one has been utilised, or even the connection made.

Related Work

Margaret Boden initially proposed novelty and value as desired criteria in computational creativity tasks (Boden 1998; 2004). Ritchie subsequently proposed a summative evaluation method by judging the product of creative systems for typicality/novelty and quality (Ritchie 2007). Colton (Colton 2008) alternatively emphasises the importance of process through assessing the presence of three criteria: *skill, imagination* and *appreciation*. Later, the FACE/IDEA models were designed to describe and capture the impact of creative acts (Colton, Charnley, and Pease 2011). The SPECS system (Jordanous 2012) was developed to resolve the need for clear and defined benchmarks across com-

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

putational creativity evaluation. SPECS evaluates systems against 14 factors identified through creativity studies. Evaluating computational creativity systems can also be undertaken via Turing-style comparison tests (Pearce and Wiggins 2001; Boden 2010), though such tests are criticised (Pease and Colton 2011).

Prompt engineering research is limited. Current work includes a six-type prompt taxonomy (Oppenlaender 2022b) and prompting design guidelines (Liu and Chilton 2022). Additionally (Oppenlaender et al. 2023) investigate perceptions of TTI generation, such as possible applications, dangers and concerns. A number of authors have explored the *skill* of prompt writing (Chang et al. 2023; Oppenlaender, Linder, and Silvennoinen 2023; McCormack et al. 2023).

Evaluating Prompt Engineering

This section offers the beginnings of a discussion for the evaluation of prompting. This discussion is in light of recent prompting research that bypasses meaningful evaluation (Chang et al. 2023), even if such evaluation is borrowed from the CC field or otherwise.

Product

Image The goals of text-to-image systems such as DALL-E is to generate images according to a given prompt. An essential sub-goal is the generation of images that properly express the creative and aesthetic aims the users expresses via the prompt. The user will likely seek to generate subjectively novel and quality images, though the achievement of this goal is contentious, which I later discuss. Despite this, the image is still an interesting object of discussion. In online communities, users share their images according to themes (sci-fi, fantasy, horror, photography, etc), where they can receive feedback or praise.

Prompt A secondary aim within TTI communities is the creation of novel and valuable *prompts*. This sub-goal is achieved sometimes, and is validated through the sharing and sale of prompts [2]. Significant value is often ascribed to the prompt as part of the "artwork" (Chang et al. 2023), however novelty and value in the prompt is entirely distinct from a novel and valuable image: though the two are commonly conflated. As with the image output, the legitimacy of prompt engineering as a skill is contended (McCormack et al. 2023), though some argue that experience with the training set, the models latent space and using particular prompt modifiers evidences a skill (Oppenlaender 2022a). Creating novel and valuable prompts relies on a novel approach to linguistic expression and traversing the latent space. An artist who is able to express a vision through the use of unexpected and surprising prompts evidences more skill than a user who is able to cycle through prompt modifiers, even if the latter produces "better" images.

[2]promptbase.com

Portfolio The ability to rapidly generate and edit high-quality images allows users to quickly build portfolios. Where it may take an artist 5 years to develop a sizeable body of work, a user could dedicate a day. The curation of an aesthetic and style within a portfolio is another way a user may exert creative control. Prominent "AI artists" cultivate a specific style, which they often mint as NFTs and try to sell.

Evaluation Boden (Boden 2004) makes the important distinction between P-creativity (novel to creator) and H-creativity (novel to culture). In the context of prompting and generation, we concerned with the production of novelty relative to its initial state of knowledge (P-creative) (Ritchie 2007). and we can relate the P-creative to the individual and community generating the prompts. Ritchie's development of 14 (later 18) criteria defines three key mappings: *novelty* and *typicality* in the intended domain and *value* of the output (Ritchie 2007; Boden 1998). Ritchie defines an inspiring set I, wherein the formal account of creativity is judged according the replication or imitation of I. Suitably novel outputs V are therefore derived from the output set O. The degree of creativity is determined by the number of novel output V produced which are not in I (Colton et al. 2002). "Fine-tuning" (Colton et al. 2002) is when systems evidence replication to a greater extent than the generation of novel high-value items. It has been proven that systems such as Stable Diffusion generate statistical amalgamations of the dataset, evidencing reconstructive memorization and imitation (Somepalli et al. 2022). This is not always easily recognised due to sheer size of the datasets. The prompt as output also evidences such limitations, many prompts that utilise guides or common modifiers are fundamentally not novel, and as such novelty and value arises in the unexpected use of language, which is arguably finite and bound by linguistic limitations (McCormack et al. 2023). Qualitatively, we can argue for novelty in output (i.e this image has not existed before), however quantitatively, it is proven that true novelty (not the imitation of) in contained TTI generation is difficult given the limitations of only rendering that which always exists (McCormack et al. 2023). However, we are also able to consider the presence of value in the form of writing the prompt if we consider the prompt as a novel creative act. We could consider the prompt process as akin to writing a series of exploratory questions.

To argue either side is to decide whether such forms of creation predicated on amalgamation, imitations, pastiche and mimicry (even possibly *unrecognisably* so) can ever represent novelty. Importantly, this does not hold for individual creative processes, only simplified prompt engineering. This exemplifies the current divide in TTI research (Chang et al. 2023; McCormack et al. 2023).

Clearly, Ritchie's criteria present a number of theoretical issues: such aesthetic measures are highly subjective and practically difficult to implement, and offer no answers for evolving generated outputs to evidence novelty without expanding the capabilities of the system beyond the inspiring set. With value and novelty contentious criteria, the IDEA ((I)terative (D)evelopment(E)xecution-(A)ppreciation) descriptive model (Colton, Charnley, and Pease 2011) offers a second path of product evaluation. The IDEA model is composed of two tasks. The first describes the stage of development, the second posits the *impact* of creation as opposed to the value metric. The IDEA model supposes an ideal audience *(i)* and quantitatively measures the impact a creative act *(A)* has on *i*. Disregarding the subjective metric simplifies many of the problems attached with evaluating prompt and image. Instead, we evaluate according to the ideal audience. In evaluation of the prompt and image as a mutually reinforcing art piece, we ideally evaluate the outputs (prompt and image) according to their proximity to each other and the dataset. Ideally this measure includes (for example) shock and subversion. Additionally, the IDEA model supposes two further simplifying solutions: ideal development process and ideal background knowledge information, which, alongside creating the ideal audience, may be as challenging as generating the creative artefact.

The application of evaluative methodologies to prompt engineering is messy at best. The above examples have aimed to show just how difficult it can be to define exactly how we think about value and novelty as requirements of creativity, especially in closed generative models. In addition, a user may rate their outputs as novel, valuable, unexpected or appealing, and therefore call themselves an artist. Indeed many in the community do. Therefore any possible evaluation of product must not *rely* on the user self-assessing, as has been done in previous studies (Chang et al. 2023), but must consider evaluation by expert users and audience. The prevalence of self-assessment and validation has only supported the criticisms levied at the communities, such as in artistic theft. However, preliminary analysis of online communities reveals a growing body of users who consider their outcomes valuable and original. This conclusion is largely premised on their reluctance to employ existing style words, artist names and over-used prompt techniques. As such, they would be an interesting place to start with this evaluation.

Process

The creative process can be broken down into a number of stages, for example preparation, incubation, illumination, and verification (Wallas 1926). Prompt engineering processes do not fundamentally differ from other creative processes, except that some stages (or tasks) are undertaken by a generative model. Much of the process is also undertaken as an iterative interaction between human user and model.

Iteration Prompt engineering has previously been broken down into two native tasks: iteration and curation. The central goal of the iteration task is to refine textual descriptions according to the previous generation in order to reach a desired image. Users must navigate and map the model's latent space via text, often times finding strange, seemingly unrelated connections or glitches. The iteration process is co-creative as both user and machine contribute to the problem solution, and should be evaluated as such. (Chang et al. 2023) found that a common goal for users was to pursue new capabilities through the creation of a specific visual language: employing words from differing domains alongside their natural vocabulary. The user's creative logic and expression is altered by the billions of text-image mappings.

Curation Users may curate an image or images through editing techniques such as inpainting, outpainting or retouching. Image synthesis models frequently fail to properly render spatial arrangements, faces or text, or simply do not achieve the goals of the prompt. Often, the creative aims of the image are reached within iteration, and curation simply resolves the expected failings of the generator. However, curation can also be a creative task. Artists may use the curation phase to exert creative agency through minimal or extensive editing, such as involving other mediums and tools, or using the generated image as inspiration or a fragment of a larger creative vision. One artist uses generative models to produce human forms, which are then painted over, another uses them to create portions of a collage[3]. The exertion of creative agency by the user is a oftentimes where value arises. Framing information - such as intention or process - are key to legitimising the final image as the result of a meaningful creative act, rather than mere generation.

Collaboration It is difficult to quantify the influence of the community on the artist. From a distance, it is possible to see how new techniques, styles and subjects disseminate, however in proximity, art appears a nebulous and interwoven world. Prompt engineering offers an unusual insight into collaborative creativity. We can call this an instance of P-creativity In the many Discord-based communities, users directly copy prompts, images, techniques, applications and ideas (Oppenlaender 2022a). This collaboration is uniquely supported by the inability to copyright AI-generated images, and the extremely low barrier to entry: anyone write prompts and contribute to the community. It is important to recognise that artists whose style, name and artworks are adopted by the prompting communities are also - unwillingly - drawn into this process.

Evaluation When we appreciate an artefact, we are appreciating both the process and the outcome (Colton 2008). We acknowledge the skill, time, dedication, knowledge and application of the artist. Our perception of how something is produced can influence our reception of the outcome (Colton 2008). This is especially applicable to prompt engineering, wherein we may misjudge or undervalue an artefact because we believe it is generated. Traditional artists have taken to posting their process - framing information - to prove their work is not generated. The digital image and prompt tell us little about the skill of the creative process and therefore user autonomy in the creative process is highly valuable. A user may have copied a prompt, utilised prompt "cheat words", or simply have stumbled upon a good output. At the same time, the user may have undergone an extensive iteration and

[3]These insights were gained from personal conversations with artists.

curation process, guided by a focused creative vision. Similarly, users can fork, train and alter TTI systems to their own specifications, which can be a creative skill in itself. Importantly, it is not fruitful nor useful to apply process evaluation to the act of simply typing a prompt - "bear in a suit" - and generating an image. We expect a user to employ and defend some creative process, artistic, linguistic, collaborative or otherwise.

To evaluate prompting processes, we must employ multiple evaluative methodologies. The first is utilised for the evaluation of mixed-initiative co-creativity, and aims to quantify the degree of use of the generated images and the quality of use within the path of creation (Yannakakis, Liapis, and Alexopoulos 2014). In this case, the goal of assessing the degree and quality of use is to conclude whether or not the generative model fosters or undermines the creativity of the user. We ideally want to understand the quality of use (understanding, subversion, evolution, exploration, re-appropriation) through asking the subsequent questions. For example, a human audience can be used to reveal the usefulness of TTI systems outside of mere generation: how useful are they in iterating through ideas? Can we identify milestones (Yannakakis, Liapis, and Alexopoulos 2014) where the user feels creatively undermined or supported? It is also important to quantify the impact that communal co-creation has on an individual user. For example, how are creative processes undermined or enhanced by community resources? How does a user seek and identify novelty and value in their product in light of the limitations of textual commands (Chang et al. 2023).

The FACE model (Colton, Charnley, and Pease 2011) captures and emphasize the importance of the process of artefact creation in a judgement of creativity. Prompt engineering and generation express multiple instances of individual generative acts. The TTI system performs creative acts of the form C^g, E^g as an executable program and subsequent execution by the user. It is also possible to argue that process evidences A^g (a local aesthetic) as images and prompts are judged according to a users given heuristics. F^g or *framing information* (natural language text that is comprehensible by people) in the form of the prompt or user-provided explanations arguably adds value by imbuing the output with meaning, and linking them to some human motivations The FACE model can be used comparatively, for example $<C^p> \; > \; <A^g> \; > \; <C^g> \; > \; <E^g>$, wherein we prioritise the method of generating concepts, in which prompting and generation may score badly, as we can argue that the generation act is imitation. It is also suggested to utilise the FACE model $CA1 = <C^g, E^g> \; < \; CA2 = <A^g, C^g, E^g>$ where the invention or choice of an aesthetic by the computer is 'more creative'. Unsurprisingly, it is difficult to apply the FACE model to a process we have categorised as co-creative. However we can apply aspects, such as prioritising the method of generation and the invention of an aesthetic in evaluating process. It would be interesting to evolve the TTI process by enabling greater agency through the generation of framing information, for example.

The final evaluative methodology is the Creative Tripod (Colton 2008). Colton argues that in order for software to be perceived as creative, it should display three behaviours: *skillful*, *appreciative*, *imaginative*. Whilst a more simplistic approach, this framework can provide insights into developing prompt engineering and generation. Any party can also contribute to the tripod (programmer, consumer and computer). If we extend this to consider the user, we can argue that a user frequently evidences skilful interaction with the system through prompts, though the skill may not yield creative or valuable results. Whilst we cannot call the TTI system "appreciative" or "imaginative", we can recognise the fine-tuned capabilities of the system to generate impressive, realistic and artistic images. An inclusion of Ritchie's criteria (Ritchie 2007) to evolving such capabilities may also yield more "creative" processes. Application of the tripod to prompt engineering is difficult, as we largely care about the co-creative relationship rather than the empty appearance of creativity.

Problems, Criticisms and Future Work

Artistic endeavour frequently manifests as divergence away from established mediums, forms, tools, techniques and subjects. Many cite the *Portrait of Edmond Belamy* and the brief popularity of NFTs as watershed moments in how artists can create art, and how customers purchase it. However this acceptance of a new suite of technologies has ignored many legal and ethical concerns. In addition, "AI art" is not always well received. Job displacement, market saturation, data laundering, copyright infringement and artistic legitimacy are only some of the issues up for debate. In addition to such concerns, it is often argued that the act of prompting does not diverge from the previous method of human-computer interaction via textual commands, and is limited to the combination or exploration of defined concepts or objects which can be expressed via natural language (McCormack et al. 2023). In this way, prompt engineering is akin to a database query. Further, it is easy to buy prompts or even generate them[4]. Further, TTI systems are dependent - even *parasitic* (McCormack et al. 2023) - on existing and new human visual data to generate 'new' images, without which outputs would devolve into pastiche.

Considering this, it is difficult to foresee widespread acceptance of prompt engineering as an *artistic* practice. Yet, it is likely that the adoption of such tools will only increase. It is important to note that the combination and expression of concepts via natural language is a foundation of human knowledge production, and undermining prompt engineering as a creative practice because of the limitations of language would undermine countless creative acts. By extension, prompt engineers are well-supported in calling their process creative yet it is interesting to consider the forms of divergence that could legitimise the process. For example, where a system is altered to provide debate or increased interactive tangibility[5] rather than mere generation. Divergence could also be realised when a user subverts the in-

[4]https://huggingface.co/succinctly/text2image-prompt-generator

[5]One artist mentioned that without two-way digital or physical interaction the process does not feel creative.

tended use of the model, exposes or alters the fundamental processes of generation.

Conclusion

This paper is a preliminary discussion of what generative deep learning can learn from a CC perspective, from the view of evaluation. I would suggest that generative deep learning has largely ignored the CC literature in system development because they do not consider creativity as a compelling aspect of generation interactions, rather focusing on developing "better" systems. I have hoped to show that CC evaluation offers a method to assessing system limitations, whilst also offering insights as to developing systems to better assist in (co-)creativity. For example, this paper has presented a number of failings in generative models such as pastiche and imitation, limited interaction and opaque process. It is also important that the CC field considers what it can learn from the mass use of deep generative models in a creative context, as these new interactions offer ripe opportunity to understanding the processes and interactions of the user with "creative" systems. This work is presented with the intention to pursue further analysis, but I have hoped to exemplify some of the connections to be made between the fields.

Acknowledgments

I would like to thank the anonymous reviewers for their kind and constructive feedback. This research was supported by AHRC BRAID programme AH/X007146/1 and the Baillie Gifford PhD scholarship at the Centre for Technomoral Futures.

References

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103(1):347–356.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Psychology Press.

Boden, M. A. 2010. The Turing test and artistic creativity. *Kybernetes* 39(3):409–413.

Chang, M.; Druga, S.; Fiannaca, A.; Vergani, P.; Kulkarni, C.; Cai, C.; and Terry, M. 2023. The Prompt Artists. arXiv:2303.12253 [cs].

Colton, S.; Pease, A.; Ritchie, G.; and Bridge, S. 2002. The Effect of Input Knowledge on Creativity. *Technical Reports of the Navy Center for Applied Research in Artificial Intelligence.*

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational Creativity Theory: The FACE and IDEA Descriptive Models. In *Proceedings of the 2nd International Conference on Computational Creativity*.

Colton, S. 2008. Creativity Versus the Perception of Creativity in Computational Systems. In *AAAI Spring Symposium: Creative Intelligent Systems*.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4:246–279. Place: Germany Publisher: Springer.

Liu, V., and Chilton, L. B. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems*, 1–23. New Orleans LA USA: ACM.

McCormack, J.; Cruz Gambardella, C.; Rajcic, N.; Krol, S. J.; Llano, M. T.; and Yang, M. 2023. Is Writing Prompts Really Making Art? In *Artificial Intelligence in Music, Sound, Art and Design: 12th International Conference, EvoMUSART 2023*, 196–211. Springer-Verlag.

Oppenlaender, J.; Visuri, A.; Paananen, V.; Linder, R.; and Silvennoinen, J. 2023. Text-to-Image Generation: Perceptions and Realities. In *Workshop on Generative AI in HCI (CHI '23')*.

Oppenlaender, J.; Linder, R.; and Silvennoinen, J. 2023. Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering. arXiv:2303.13534.

Oppenlaender, J. 2022a. The Creativity of Text-to-Image Generation. In *25th International Academic Mindtrek conference*, 192–202. arXiv:2206.02904.

Oppenlaender, J. 2022b. A Taxonomy of Prompt Modifiers for Text-To-Image Generation. arXiv:2204.13988.

Pearce, M., and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. *Proceedings of Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*.

Pease, A., and Colton, S. 2011. On Impact and Evaluation in Computational Creativity: A Discussion of the Turing Test and an Alternative Proposal. In *AISB 2011: Computing and Philosophy*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs].

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines* 17(1):67–99.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs].

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2022. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. arXiv:2212.03860 [cs].

Wallas, G. 1926. *The Art of Thought*. Harcourt, Brace.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *Proceedings of the 9th conference on the foundations of digital games*.

AI as *other*: An art-as-research approach to generative AI art practice

Suk Kyoung Choi
School of Interactive Arts and Technology
Simon Fraser University
Surrey, BC, Canada
choisukc@sfu.ca

Steve DiPaola
School of Interactive Arts and Technology
Simon Fraser University
Surrey, BC, Canada
sdipaola@sfu.ca

Abstract

This Art-as-Research explores emergent processes that develop between creative practitioners and artificially intelligent (AI) technology when an AI system is positioned as ontological *other* that an artist works *with* to produce an image. The authors, as artists and early adopters of AI image synthesis, create aesthetic artefacts and investigate the artistic process used to visualize and conceptualize creative praxis in this new media, critically examining how generative AI systems impinge on and enhance creative freedom, mediating the essential relation between self and practice. This process dynamic is employed as a phenomenological probe into AI generative art. The authors examine how artistic intention is reshaped by algorithmic transformation and re-presentation to question what is preserved, nurtured, lost, or irrevocably altered in the interplay of the autographic and the algorithmic. The study finds that *neural media*, as the authors term it, is a reflection of the ambiguous mediation of input and redirection of intention, motivating an *anticipatory aesthetics*. The non-deterministic processes in generative AI systems create an external perturbation of the artist's innate expression of the mental image. This disruption provides an ambiguous computational "other" in the artist's practice environment, expanding the field of interactive potentiality and augmenting embodied intentionality.

Introduction

In this paper we explore the experiential relations between the situated artist-researcher and artificially intelligent technology conceived of as distributed *ontological other*, a virtual collaborator co-involved in a dynamic interaction that the artist-researcher works *with* to produce an artwork.

We look at a supposed support technology (generative AI art software) and observe how it goes beyond support of intention and becomes a mediating influence embedded in the creative process. Our motivating concern is to show both the enhancement and limitation of a technology's shaping of us, by asking how generative AI systems impinge on and enhance the artist-researcher's creative freedom. It is this essential relation between self (our subjective emotive and intuitive being) and practice (our objective actions in the world situating and developing that being) that we employ as a phenomenological probe into AI mediated art process and the formulation of what we have called *anticipatory aesthetic praxis* (Choi 2021). In our research we ask how the intentions of the artist are reshaped by algorithmic mediation, a re-presentation that questions what is preserved or nurtured and what is lost or irrevocably altered in the agonistic polarity and interplay of the autographic (production from the artist's "hand" and "mind") and the algorithmic (production through computational systems).

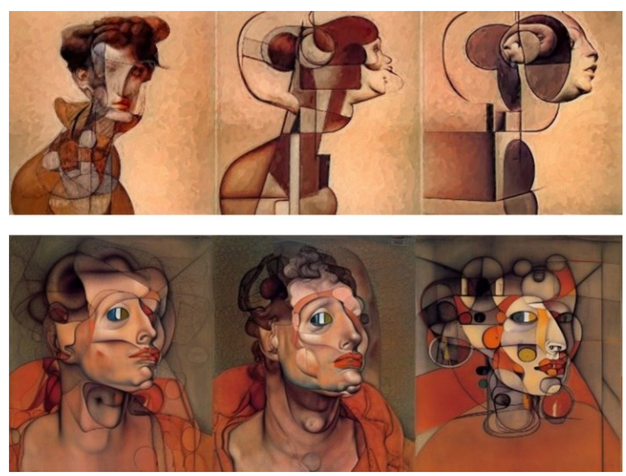


Figure 1: Two examples demonstrating movement through concept or style. Diffusion models allow for experimentation through complex association along any axis of image-concept in latent space. (Image © S. DiPaola 2023).

As established computational media artists, we explore newer technologies associated with machine learning, working for years with Deep Dream systems, then Neural Style, then GANs and more recently diffusion-based generative systems. Our initial approach to newer diffusion-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

based "text to image" and "image to image" systems (that use NLP text prompts and large, trained datasets) started by attempting to mimic traditional autographic practices such as painting or sculpture. This alignment is perhaps to be expected as we approached the new experience based on a pre-existing set of assumptions about how one might proceed in a creative image space. The "errors of innocence" in this approach turned to an advantage, as we quickly ran into valuable instructive "perturbation" of our anticipations that were richly revealing of our entrance assumptions. An initial realization was that *neural media* as we have termed it (Choi, DiPaola, and Töyrylä 2021; but see also Choi 2018) reflects the ambiguous mediation of input and redirection of intention inherent in the human-AI relation. This embedded ambiguity drives the evolving composition forward in iterating cycles of divergence and convergence.

This process dynamic was initially frustrating but soon became a state of "serendipitous release" affording creative opportunities we would not have arrived at through the methods we assumed we might continue working but that were biased by those prior assumptions. An anthropomorphic sense of "playing against another perception" emerges though the growing awareness of a mediation taking place that is not exactly controlled but simultaneously in no way presents a feeling of randomness. This form of affective-technical interaction suggests that if computational output inspires reflection in interacting humans, then we have a technology that is already intersubjectively improvisational by nature: a complementarity of improvisational exchange emerges through the creative process where the artist does not control but only suggests (Figure 1). We found that the embedded "alterity" relation (Ihde, 1990) implicit within artificial intelligence research promotes creative practices that are expressly intertextual, simultaneously subjective and distributed, taking place at the multimodal interstice of image, text, and code. The singular source referent recedes to the background, and traces of the source images and textual input appear throughout a sequence of generated images but enter an ambiguous space of latency where representation and abstraction define horizonal limits to the *potentiality* of the image but do not enter into any explicit immediacy with the content of the output; instead, an imaginative, immanent image is suggested and the artist takes on a curatorial role, allowing some streams to proceed while terminating or modifying others speculatively. The artist becomes a finder of regions of cultural attraction (Buskell 2017) more than the author of singular experience. The artist-as-researcher examines the process-motivated transformative and ecological sources of this convergence on points of latent multimodal space through a phenomenology of AI mediated manifestation of the imaginary image.

The autographic and the algorithmic

In this research the anticipatory relation of the interacting artist with the digital aesthetic artefact is speculatively positioned as an externally mediated affective process accepting expressive actions and returning modulated reflections of creative intention. We draw from the phenomenology of this interaction that generative AI systems can be perceived as "life-like" precisely because interaction with them is non-deterministic and poses a distributed perturbation of the artist's naturalistic/autographic seeking of the mental image. This disruption presents an ambiguous computational "other" in the artist's otherwise familiar praxis environment. A phenomenology of existential distinction is therefore centralized in AI mediated aesthetic practice. Generative AI widens the environment of creative practice beyond the strictly intentional as there is always some undisclosed element that plays into the interaction which cannot be directly interacted *with*. This is unlike traditional autographic media such as oil painting where tactile interaction (of brush to canvas, or the multimodality of the scent of paint and the warmth of a beautiful day) is more immediately engaged with and embodied into praxis knowledge. The AI latent space is thus abstracted from lived experience but affords an expanded field of anticipatory potentiality *augmenting* embodied intentionality *through* disruption of situated expectation. The resulting anthropomorphic overlay of an implied "theory of mind" in the interaction with AI technology motivates an intentional stance (Dennett 1989) toward the tool and implies that artistic expression as an evolving process of self-apprehension leaves in its wake a data trace—a praxis narrative of affective intent in the multimodal ecology of creative practice—from which AI might learn about and reflexively extend human anticipatory acts. Although it is generally acknowledged that human-centered practices are extensively multimodal by nature—as evidenced for instance in the rising awareness of the essentiality of rich data in medical practice (Acosta et al. 2022)—there so far has been little development of robust frameworks of affect-oriented multimodality in AI network architecture. Recent work by Google Research (2023) on PaLM-E, a large language model coupled with an advanced vision model, attempts to demonstrate the potential of situated "embodiment" in AI robotics: Experiments show that the PaLM-E system model is capable of developing untrained viable real-world behaviors in complex tasks. Rich computational multimodality will be necessary to model and support human level causal behaviors and AI generative art praxis is an ideal testing ground for studies of affective response to human-centered generative technology deployed in an environment that is situated, persevering, non-destructive, and critically and aesthetically multimodal.

In traditional painting or drawing, the autographic artefact represents a set of past assumptions, informing the transactional nature of embodiment where the function of metaphor is to guide the accumulation of sensorimotor acuity and tacit knowledge, rather than establish schema for the manufacture of objects. The object of art obsolesces at the project's completion as what the artist was looking for has been absorbed into being, encoded into future anticipatory projections while simultaneously released from concern. However, artificial intelligence development has obscured this distinction between imagination (potentiality) and virtuality (artificiality), offering in return a conjoined *hyperobject* (Morton 2013) composed of an ambiguous and inseparable blending of technological and environmental epistemologies. We suggest that this hyperobject—an entity that is present but never completed/situated—constitutes the creative and ethical imperative of the *Anthropocene*, the perhaps limited "age of humans" (Crutzen and Stoermer 2000) that may be drawing to a close just as our most advanced technology emerges (Colebrook 2014). If we are not willing to question the horizonal extents of the post-human, then we have already opened the Trojan horse (or Pandora's box – pick your metaphor) of an AI mediated and predefined future. We conceive of this aesthetic hyperobject as a metaphorical warning myth and humanist critique of the ethical imperative we find ourselves in today with the AI entanglement of the virtual and physical environments, the one rising the other falling, but now conjoined and inseparable. We propose that "perspectival affordance" in an AI generative ecology of functional and embodied relations in the creative praxis of neural media may be instructively engaged with as reflective of the problematization of an unacknowledged ethics of the Anthropocene. The intent here is to sketch out a set of conceptual relations encountered in the phenomenology of neural media so that further analysis of the relation between embodied cognition and its AI representation might be grounded on more authentically experiential frameworks.

We use art-as-research to both create artefacts and investigate the process we engage with to understand and conceptualize praxis in this emerging media environment. Art-as-research (Barone and Eisner 2012; Biggs and Karlsson 2010; Klein 2017) is a field of study that is growing along with the realization that "big data" alone may not be enough, or the right kind of data, to teach creativity to artificial intelligence or even to train statistical inference engines (Mitchell 2019; 2020; but see also Shilo, et al. 2020 for similar issues raised in healthcare). The establishment of "point of view" of situated cognition is central to practice-based research where iterative granular interaction with an emerging artefact of expression may only be perceivable at close range by an involved but detached observer. We argue, in this contemporary explosion of AI advancement, for the possibility of *metaphoric alignment* of the subjectivity of art-as-research with the objectivity of intelligent technology development. The *metonymic* sources of AI mediated *affect* are only minimally present (if at all) in mediated connectivity because the immanent potential of any "intelligence" is beyond the event horizon of another intelligence. We simply do not see our own bias to begin with (Greenwald and Krieger 2006), so it is virtually impossible without critical reflection to see the extension and mediation of that same bias by external technologies.

Therefore, in our investigations we position the computational *apparatus* (Flusser 1984) as "other," speculatively adopting an intersubjective theory of mind that is presumed to originate from the network of programmers, engineers, and entrepreneurs that have already left their mark in the depths of the black box, but which may present an emergent gestalt intelligence beyond what can be known from the outset. As Ranulph Glanville has observed, "inside every white box are two black boxes trying to get out" (Glanville 1982), meaning that the description (observation) and the model (implementation), transparent to themselves, are opaque to each other (Figure 2).

Figure 2. A grouping of images from the same region of latent space. The intersection of several aesthetic vectors reveals a diverse region of related affective stimuli that is not necessarily transparent to the interacting artist. (Image © S. DiPaola 2023).

Discussion

From the cognitive framework set out in this research several questions and findings are identified:

1. What is the existential nature of the emergence of the aesthetic mental image in a praxis of artificially intelligent image synthesis?

We found through immediate subjective interaction apart from all but the most basic operational scripting that artistic process "loses touch" by which we mean that compositional intention is distanced from tactile interaction with the body and refocused on the intellectual, and in some displaced way a transformed-emotive, interaction. The technology thus—and rather curiously—reinforces by design the Cartesian metaphor of the separation of mind (as "software") and body (as "hardware"), a reflection of the machine metaphor adopted early in the development of computation and still prevalent today (Searle 1990).

2. How do artificially intelligent image synthesis technologies mediate the embodied intention of the artist in manifesting the tacit image?

After working extensively through the experimental creation of many AI images while concurrently maintaining our other more traditional art practices, we observe that there is an agonistic divide between the autographic and algorithmic. The two may be exchanged but never lose their individual mediation. The neural media artist is constantly in a state of translation between media rather than at play with an emergent (blended) third state. This is not a restriction as much as an apparently ontological feature of the variant latent spaces that emerge from the two media ecologies. There appears to be a relation of space and time that is divergently emphasized; autographic painting weighs toward space, algorithmic painting toward time. This is reflected in the constraints of the body as foundational to the former and velocity of information as definitive of the latter. Moreover, "speed" of information is associated with interconnectivity as information density promotes an intertextual hermeneutics where creative agents sample from, remix, and recontribute to the global networked data flow (Jenkins 2006) in a transient flux of non-linear association emphasizing the "systemic 'malleability' of digital information" (Rigney 2010, p. 112). This "malleability" however exhibits a certain polarity, that is, space is drawn into time more than time is drawn into space. So, digital information captures the autographic through data sampling, extending and augmenting its presence, whereas autographic expression is limited to the physical dimensionality of the medium and some specified partition of time allotted to the interaction.

3. Given what has been revealed, what kind of conjoined entity is the algorithmic aesthetic artefact?

We find that situatedness takes an ontological shift to an alternate computational aesthetic. When properties become distributed across a network of perpetually reconfiguring relations, and the objects of attention themselves are virtual, transient, and simultaneously ubiquitous, then a new *anticipatory aesthetics* that is more computational than singularly human emerges. That emergence poses an affective relation with the virtual artefact that is as much (if not more so) temporal than locative and physical, and the aesthetic is then extended across time, widening the existential horizon of the aesthetic experience. We therefore observe that situated accounts emerge from *transactional selves*. The algorithmic artefact appears as the trace of resonance in a latent space of possibility, a multimodal intersection of ongoing processes rather than a constitution of situated materials.

4. Why does algorithmic art need an artist?

This question forewarns that the Anthropocene may become mediated by some higher form of semi-intelligence and humans will be "none the wiser" –a situation that could too easily lead to a legislated devolution of consciousness and which may already be entering the historical record (Crawford 2021; Harari 2023). In response, the pragmatic critique draws from a certain amount of skepticism that asks whether science has yet been able to save us from our seeming wish to destroy our home world. Therefore, "machine art" for the foreseeable future is likely to consist of human-directed algorithmic manipulation of data, or automated routines running on fallible hardware producing virtually endless variations on the same piece of code they started from. Despite the fear and fandom surrounding the technical "singularity" (Vinge 1973), and importantly in terms of the themes of our work, an ethics of algorithmic autonomy is centralized because it is still existentially horizonal, that is, we still have time to learn to perceive it from a distance. Here we might coin the term "computational subjectivity" meaning not to suggest that the machine is likely to start offering considered critique back to its human "collaborators" but rather to suggest that subjectivity may be studied in a new way when reflected through a technology which mediates the expression of that subject in ways that are reflective of intent, yet which may be opaquely shifted in unpredictable directions. This shift affords a phenomenology from which we may learn about our subjective bias and probe the black box in a tactical engagement with AI media that "mobilizes AI's emergent capabilities for interrogating, exposing, problematizing, and challenging the aesthetic, ideological, or technological frameworks driving the commodification and propertization of creative expression" (Zeilinger 2021, p. 27). Or, as Marcus du Sautoy puts it "machines might ultimately help us, as humans, to behave less like machines" (Du Sautoy 2019).

But reflective AI requires reflective humans, and reflectivity implies a process of deep introspection that Merleau-Ponty calls hyper-reflection (Toadvine 2014), a reflection that is not a "temporal exercise" but a reflection *on* reflection (Daly 2016, p. 294, 295), a deepening awareness of one's self-looking, from "outside" as it were, an introspection that prioritizes the subject through its self-removal. This is the importance of subjective studies, for when the subject becomes objectified—through an insistence on generalization as reinforced by algorithmic media—then we are pre-defined by a set of externally mediated cultural controls. Is this the AI we want to live with? What might algorithmic introspection look like—the deep reflection of the technically embodied subject, an *apparatus* of self-awareness? We want to suggest these questions insist on an empathic resonance with technology conceived of as self *as* other.

Conclusion

For better or worse we are now irrevocably entangled with technologies that insert highly abstract and invisible codes into every gesture. Tactility, if not lost, is taking on different relations with the body-mind, reprogramming embodiment with every communication. Martin Zeilinger proposes that this "post-human agential assemblage" might be turned back on itself to offer a strategic disruption of the systemic assumptions of ownership that have instituted the tools of its arising and therefore "[t]he emergence of the posthumanist assemblage in which the agency for expression, creativity, or authorship might be distributed across multiple entities (human and non-human alike) hinges on a radical rethinking of what property means and how it operates, what we mean by cultural ownership, by creativity, by calling something a creative expression" (Zeilinger 2021, p. 173). This distributed subjectivity, we argue, if it is to survive as creative human spirit in the AI Anthropocene, must couple deep reflection with an anticipatory aesthetics of inter-subjectivity, so that we might perceive the existential horizons collapsing around us as we imagine beings of endless virtuality. A radical rethinking of property is a radical rethinking of self.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. 2022. Multimodal biomedical AI. *Nature Medicine*, 28(9). doi:10.1038/s41591-022-01981-2

Barone, T., & Eisner, E. W. 2012. *Arts Based Research*. doi:10.4135/9781452230627

Biggs, M., & Karlsson, H. 2010. *The Routledge Companion to Research in the Arts*. Routledge.

Buskell, A. 2017. What are cultural attractors? *Biology & Philosophy*, 32(3): 377–394. doi:10.1007/s10539-017-9570-6

Choi, SK. 2018. Guess, check and fix: a phenomenology of improvisation in 'neural' painting. *Digital Creativity*, 29(1), 96–114. doi:10.1080/14626268.2018.1423995

Choi, SK. 2021. What happened to the subject? Mediated anticipation in neural painting. *Technoetic Arts*, 19(3) (Themed Issue: 'On Modes of Participation'): 301–320. doi:10.1386/tear_00071_1

Choi, SK., DiPaola, S., Töyrylä, H. 2021. Artistic Style Meets Artificial Intelligence. *Journal of Perceptual Imaging*, 4(2). doi:10.2352/J.Percept.Imaging.2021.4.2.020501

Colebrook, Claire. 2014. *Death of the PostHuman: Essays on Extinction*, Vol. 1. Open Humanites Press.

Crawford, K. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Crutzen, P. J., & Stoermer, E. F. 2000. The Anthropocene. *The International Geosphere–Biosphere Programme (IGBP) Newsletter 41*. 17–18.

Daly, A. 2016. *Merleau-Ponty and the Ethics of Intersubjectivity*. London: Palgrave Macmillan.

Dennett, D. C. 1989. *The Intentional Stance*. MIT Press.

Flusser, V. 1984. *Towards a Philosophy of Photography*. Reaktion Books.

Glanville, R. 1982. Inside Every White Box There are Two Black Boxes Trying to Get Out. *Behavioral Science*, 27(1): 1–11.

Google Research. 2023, March 10. PaLM-E: An embodied multimodal language model. https://ai.googleblog.com/2023/03/palm-e-embodied-multimodal-language.html

Greenwald, A. G., & Krieger, L. H. 2006. Implicit Bias: Scientific Foundations. *California Law Review*, 94(4): 945–967. doi:10.2307/20439056

Harari, Y. N. 2023, April 28. Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist*. https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation

Ihde, D. 1990. *Technology and the Lifeworld: From Garden to Earth*. Bloomington and Indianapolis: Indiana University Press.

Jenkins, H. 2006. *Convergence Culture: Where Old and New Media Collide*. New York and London: New York University Press.

Klein, J. 2017. What is artistic research? *Journal for Artistic Research*. doi:10.22501/jarnet.0004

Mitchell, M. 2019. Artificial Intelligence Hits the Barrier of Meaning. *Information*, 10(2). doi:10.3390/info10020051

Mitchell, M. 2020. On Crashing the Barrier of Meaning in Artificial Intelligence. *AI Magazine*, 41(2). doi:10.1609/aimag.v41i2.5259

Morton, T. 2013. *Hyperobjects: Philosophy and Ecology after the End of the World*. University of Minnesota Press.

Rigney, A. 2010. When The Monograph Is No Longer The Medium: Historical Narrative In The Online Age. *History and Theory*, 49(4): 100–117.

Sautoy, M. du. 2019. *The Creativity Code: Art and Innovation in the Age of AI*. Cambridge MA: Belknap, Harvard University Press.

Searle, J. R. 1990. Cognitive Science and the Computer Metaphor. In B. Göranzon B.; and Florin M., eds., *Artificial Intelligence, Culture and Language: On Education and Work*. Springer. 23–34. doi:10.1007/978-1-4471-1729-2_4

Shilo, S., Rossman, H., & Segal, E. 2020. Axes of a revolution: Challenges and promises of big data in healthcare. *Nature Medicine*, 26(1). doi:10.1038/s41591-019-0727-5

Toadvine, T. 2014. Phenomenology and "hyper-reflection." In Diprose, R.; and Reynolds, J., eds., *Merleau-ponty: Key concepts*. Taylor & Francis Group. 17–29

Vinge, V. 1973. The Coming Technological Singularity: How to Survive in the Post-Human Era. https://ntrs.nasa.gov/search.jsp?R=19940022856

Zeilinger, M. 2021. *Tactical Entanglements*. Lüneburg: meson press.

Fostering Mental Well-Being through Creative Interaction: An Assessment of SOVIA

Lauryn Gayhardt*, **Margareta Ackerman***, **Lee Cheatley**+
*Department of Computer Science & Engineering
Santa Clara University, Santa Clara, CA
+Lloyds Banking Group
Edinburgh, Scotland
{lgayhardt,mackerman}@scu.edu, leecheatley@gmail.com

Abstract

Therapeutic Computational Creativity is an emerging domain that challenges us to explore applications of Computational Creativity systems to mental health and wellness. This includes the assessment of creative systems for their ability to support well-being. In this paper, we assess this potential in the co-creative system SOVIA, which engages users in a creative interactive experience with Monet's paintings. We conduct a user study followed by thematic analysis to ascertain SOVIA's value for mental well-being.

Introduction

Recent years brought awareness to the importance of mental health. The COVID-19 pandemic came with a substantial mental health toll, making it more urgent than ever to find affordable ways to help people maintain mental health and wellness. Therapeutic Computational Creativity (TCC) is a emerging field within Computational Creativity (CC) that overlaps human computer interaction, art therapy, and psychology. While TCC does not aim to replace classical therapists, it can offer benefits with therapeutic endeavors (Pease et al. 2022).

Previous work in TCC focused on bereavement. One study explored current reminiscence practices and receptiveness to CC related tools with the bereaved (Cheatley, Moncur, and Pease 2019). The following year, a user study was used to analyze ALYSIA (Cheatley et al. 2020), a co-creative songwriting machine, to assess its utility in the bereavement process. Data from the study was analyzed using *thematic analysis*, a technique that allows researchers to identify patterns in data by discovering recurring themes, allowing them to examine common experiences and meaning throughout a group of participants (Braun and Clarke 2012). The study found that ALYSIA supports self-expression, as well as helps users reminisce and gain awareness of their feelings.

In this paper, we focus on the wellness potential of SOVIA (Gayhardt and Ackerman 2021), a co-creative machines that places most of the effort on the machine agent, while giving the user a simple and enjoyable experience that deepens engagement with Monet's landscapes. A mixture of music with realistic and associated sounds creates an experience that mimics realistic elements in the art, while reflecting the gentle artistic reinterpretation of those objects through sound. As the user moves their mouse across Monet's landscapes, they hear soundscapes that represent the portion of the painting that they are currently focusing on. This results in an active experience that differs substantially from engaging with the art at a purely visual level, and helps the user "step into the art."

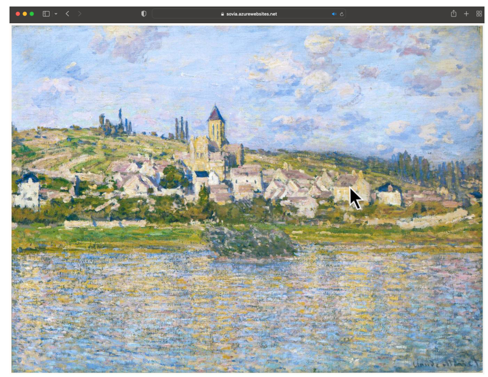


Figure 1: SOVIA offers a creative interactive experience that brings Monet's art to life. The user can hear lively conversation when the mouse is hovered over a building, cicadas with distant birds when over grass, the wind blowing when hovering over the sky, etc. A soft backing track connects the sound corresponding to the current position in the painting. In this example, Monet's Vétheuil (1879) is displayed on screen.

SOVIA works by utilizing computer vision methods to discover objects in Monet's artwork (ex. water, sky, building, tree, etc) and associating those elements with a variety of pseudo-randomly selected sounds. Whenever a user hovers over a particular object (ex. tree), a sound corresponding to this object (ex. birds chirping) is mixed with the underlying music (see Figure 2 for an illustration). Since the user directs the sound through mouse movements, combined with the pseudo-random selection of sounds, the musical dimension of the experience is co-created by the user and machine in real time. SOVIA may be accessed here [1]:

[1] To interact with SOVIA, click on the painting after it loads (may take several seconds) to start playing the sounds, then move

Proceedings of the 14th International
Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

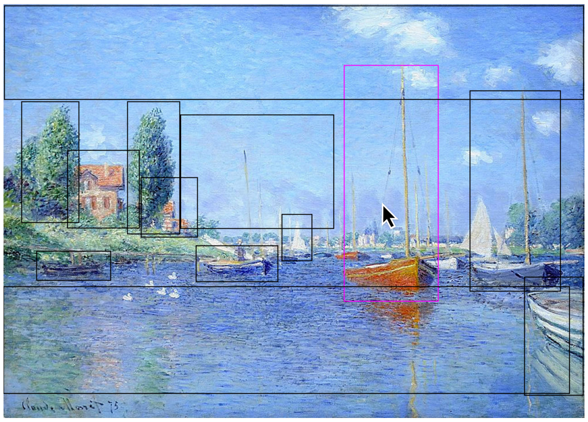


Figure 2: The bounding boxes detected using computer vision are labeled and mapped to a set of sounds, one of which is played when the cursor enters the box. The painting is of Red Boats, Argenteuil (1875) by Claude Monet. Photo Credit: WikiArt

https://sovia.azurewebsites.net/ and a video demo can be found here: https://youtu.be/XMMMBeukhb4.

In our analysis of SOVIA's potential for wellness, our focus here shifts outside of bereavement, interviewing people from the general population to assess the value of SOVIA in a broader mental wellness context. In this study, we utilize open-ended interviews with thematic analysis. First we will go over the study's methodology, then detail the results and discovered themes, and conclude with a discussion of the findings and future work.

Experimental Setup

The overarching goal of this study is to evaluate the potential uses of SOVIA for mental wellness. The study took place over zoom where participants were introduced to SOVIA by verbal instructions and a live demonstration where they were shown how to navigate the system and how to discover new images. Afterwards, participants were asked to engage with SOVIA directly and to view at least three different paintings, but encouraged to look at as many as they wanted.

To allow the users their own (unbiased) experience of SOVIA, the researcher muted their audio and video while the participant was using the system. The researcher would not interrupt unless the participant had a question, was done using SOVIA, or 20 minutes had passed. Afterwards, a semi-structured interview was conducted with each participant to explore their reaction to and experience of SOVIA. The majority of questions asked were open-ended, for example, "How did SOVIA make you feel?" and "How would you describe your experience using SOVIA?" Questions that were not open ended had follow up questions so the participant could elaborate.

your mouse around the painting and hear how the soundscape reacts to your exploration of the art. Refresh to get a new painting.

Recruitment and Participants

Participants were recruited through snowball sampling and the research participation tool at Santa Clara University's psychology department, SONA. The only requirements was that participants had to be over the age of 18, speak English fluently, and have access to a computer and the internet. The 11 participants were anonymized with the assignments of P1-P11. Eight identified as female and three as male, P5-P7. Three, P1-P3, of the participants were career professionals in the age group of 51-57. They had different levels of educational background: P1 had a Bachelor's, P2 had a Master's, and P3 had a PhD. While the others were college students working on their bachelor's degree of varying majors between the ages of 18-25.

Artistic Affinity

Participants were asked background information about their interest in art, to determine if this affected experience with SOVIA. Most of the participants had a regular appreciation for art; that is, they enjoy art, but do not go out of their way to view it, and might currently participate in art casually through doodling, coloring, etc. However, some participants were more enthusiastic about art. P7 makes art while working on game development, as in 3D modeling and textures for world development. While P10 is an active hobbyist who enjoys drawing, painting, and pointillism art. Additionally, P4 was the most passionate about art, as she enjoys going to view art physically (international museums) and is often aware of local art exhibitions. She is also an active hobbyist who enjoys painting, photography, and cinematography. P9 seemed the least interested in art out of everyone, and admitted that she didn't care for it growing up. However, she now appreciates it more through her interest in makeup art. No participants had an expert level of interest in art, that is, none studied art academically or did art professionally.

Ethics

This study was approved by Santa Clara University's Review Board before data collection took place.

Results

Five major themes were identified in the user interviews: Calm, surprise, emotional association, control, and curiosity. We detail each below. Verbatim statements were kept mostly untouched except for the omission of the filler word "like", as it makes the statements easier to read.

Calm

All of the participants felt a sense of calm, peace or relaxation when using SOVIA. One of the first reactions many participants had when asked how they felt was this feeling of calm, "I felt.. very serene and I felt very calm and each photo is kind of like a different experience."(P4) "I think it's very calm like the picture itself is very beautiful and simple calming but then you add music, on top of it and I kind of feel it comes to life, a little bit like you can kind of sense the mood of whatever's going on"(P11) "I would say relaxed...

I guess, I could say happy because it, it was a cool experience and that I've never done before." (P9) "if you close your eyes, you can almost feel like you were in wherever the big painting was set in. It was just kind of peaceful."(P6)

Participants mentioned that they think SOVIA could be used as a de-stressor. One participant even had a real-time experience being soothed by using the system. P11 is a student who was stressed about school and upcoming finals and was thankful for using it, "Well actually I think that it kind of was nice, I mean it really did calm me down because, this is week 10 is very stressful.. It was... like just for a second it was just like okay chill... I feel kind of good right now."

When asked if they would use SOVIA again and when/why many mentioned that they could see themselves using it to break up events that can be tiring or induce stress. P3 said that they would see themselves using SOVIA in between meetings at work. P2, a speech and language therapist, works with preschoolers to 6th graders also mentioned that they could use it to break up work activities, "I think that's kind of a neat way to kind of to immerse yourself into it a little bit and...just gonna be a little relaxing break maybe at work between groups [of kids]". P4 voiced how she could see herself using it as a break or before a stressful event, "If I'm stressed out and I feel like also [I could use] it in between homework assignment is kind of like a break, or before a test to kind of get my nerves down or something like that."

P6 thought SOVIA could be used as "A tool to relieve anxiety... Kinda like get you in a better mindset to fix ... or get through whatever you're anxious about." P10 also mentioned that they could see SOVIA helping someone decompress if they are experiencing anxiety, "I felt that perhaps it's sort of like maybe an individual is going through a lot, and they have a lot on their mind. And so, something like this can kind of let them decompress . . . focus more what they're seeing and what they're hearing. So I think it's kind of like when you have a panic attack or anxiety something like that, and then they tell you to list what you see or something in the room."(P10)

Some could see using SOVIA for meditation. P4 thought it could be used as a before bed meditation, while P5 saw it being useful for meditations to be present. "I think a different form of meditation but not really closing your eyes, but being present in the moment and just having your headphones in and just listening ... to the day sounds like practicing breathing." (P5)

Even those who felt that SOVIA wouldn't be their first choice for de-stressing, felt that it could be helpful under certain circumstances. "I don't know if it'd be the first thing that I go to for a de-stressor but I do think if needed, I could do that ... So maybe a before bed meditation type of thing."(P9). P10 shared that their primary choice for calming down involves going outdoors, but that they could rely on SOVIA if going outside was not an option. "I think it's nice... if I want to distress, I guess, maybe I would use SOVIA if it's late at night I can't really go outside".

Emotional Association

Using SOVIA reminded some participants of their lived experiences. Whether their experience was positive or negative affected their reception of certain sounds and visuals, and aided in recreating the feelings they had from that memory. For some users, it brought a sense of nostalgia, "I really like the birds chirping I feel like that's an association I have with like summer and good times, so that's how it made me feel like calm like at peace" (P8).

Other participants were reminded of an activity, "Some of the paintings just listening to the trees kind of reminds me of when I go on hikes. Or if I'm in my backyard and I just really need to de-stress. . . .When I go on hikes it's when I have free time and it's basically moments and times when I don't really need to think about anything too deeply I can just let myself, be a little bit free now and I I don't have to be constantly thinking about worries my problems and so." (P10) "I do meditation and, many of the the sounds I heard... from the art form really kind of remind me of my meditation" P5.

Not all associations were positive. P7 didn't like the sky sounds "The wind makes me a little bit uneased. . . wind just doesn't make me very happy". It reminds him of how he feels when in the wind "... whenever it's windy I get cold and stuff blows away and it's just hard to walk. And yeah something that's just like bad feelings that I've had in the past... it's like the worst thing ever"(P7).

Control

Many participants liked the sense of control they felt. They liked that they could choose the sounds they wanted to hear and when. For some, the aspect of control was a defining factor in their enjoyment of SOVIA. P7 details his experiences using SOVIA and where he chose to move their mouse. "I was able to go down into, the jungle area and then there was those birds... Then there was a little town and there's people talking ... [It was] very relaxing it was nice because, I could control what sounds I want I feel like I was like moving around in the picture, even though, it was just the mouse moving."

P7 also mentioned earlier in the interview that when it comes to game development he likes to do all aspects from coding, art, and design because he enjoys the sense of control. P6 also liked that he could pick what he heard "I like [that] you can kind of choose which sounds you heard in the painting. You could be next to the river and hear the water, or you can go into the air and hear the wind".

P4 liked the power she felt when using SOVIA and how she could create her own experience which enabled her to be creative, "feel like I could be creative as well, I felt like I had a lot of power and creating my own separate life. [It's like] I could kind of step back from reality for a second". While P3 compared SOVIA to the Calm app[2]; because it also uses nature sounds. However they preferred being able to control what they heard and they enjoyed that more.

[2]The Calm app consists of a variety of meditations, music, soundscape, etc. designed to aid with sleep and relaxation.

Curiosity and Surprised

At first thought, P3 said she felt relaxed after using SOVIA. When asked if relaxing is the only way she would describe the experience P3 said "I would put calming, I would put engaging it's definitely kind of elicited curiosity." P3 felt that it was more engaging because you are "connecting the sounds with the details and making you explore, so that you can hear new things... I felt like it made me notice the details of the painting more". P1 also mentioned that it made them curious, "it was a calming effect, calming and curious I guess... it inspired me to touch different parts and and look at different parts of the painting and just curious like how does that, how does this work?"

As P11 reflected more on how SOVIA made her feel, she realized it actually made her very curious, "Like just really calm and relaxed and curious .. actually curious ... and some of the sounds weren't what I was expecting so it was really cool." She later described, "... I hovered over her house and on the previous one, there was people talking and then in the next painting... one of them was silent there's no talking... It had more movement and wind, and all this stuff, so I think...it gets me curious like oh, what is going to happen next, what is this going to sound like that's not what I expected yeah." (P11).

Participants said that they felt moments of surprise when using SOVIA. P10 thought the unexpected was interesting and provided a unique point of view. "I think it's sort of like an interesting thing perhaps that's not the sound I expected to come from. Whatever I hovered over I didn't expect to hear chimes for flowers, but I think that's an interesting thing and that could sort of be in itself kind of a means of art like you know you overlay certain sounds over something and that's someone's unique perspective that perhaps you didn't consider." (P10).

Another participant also was surprised when they heard certain sounds "I just heard the voices I was kind of surprised because I was like Oh, I thought it was just like instrumental you know, but once I heard the voices that just like caught me off guard, but it wasn't really scary it's just like something surprising"(P8) Others expressed moments of excitement, "I was kind of excited to like see. Oooo what if I tap here, if I tap there."(P2).

Discussion and Conclusions

In this paper, we examined the impact of SOVIA on human well-being through semi-structured interviews and thematic analysis. The overarching theme across all interviews is that using SOVIA gives rise to a sense of calm and relaxation. For some, using SOVIA reminded them of past experiences. A number of participants noticed they were curious, while others were surprised at moments, which fueled their curiosity.

Participants were asked to share potential applications of SOVIA. A plurality of them mentioned that they think it could be used in schools to help kids be more engaged with art. One participant, unprompted, revealed "I'm a school counselor and I was thinking this could be an interesting application to do with students." (P1) Some felt that it could help people with disabilities, such as ADHD, stay engaged when looking at art. Most also voiced that it could be used to de-stress or make one calm, whether it is between school/meetings, before bedtime, or other scenarios.

SOVIA appears to be a system of connection and transition, whether it is used emotionally or socially. Participants felt a deeper connection to themselves and thought others would benefit from its usage prior to class or therapy. Using SOVIA before counseling could open the mind for lower conflict engagement before couples or adolescent therapy, or as an ice breaker to meeting a new therapist. Use before yoga or massage could deepen the experience and render more medical benefit from those activities.

Given the themes detected, it will be interesting to evaluate the impact of using SOVIA with specific populations. Stroke survivors can experience anxiety, irritability, loss of memories, among other challenges[3]. They can also experience motor apraxia. SOVIA may have the potential to aid stroke survivors, by helping them restore a sense of balance when needed. It could also assist in recovering motor function of the left/right side of the brain since using SOVIA requires using a mouse or touch as the user navigates the artwork.

This was an early study of the efficacy of SOVIA in a therapeutic context. For future work, we would like to conduct a study on a larger number of participants, carried over a longer period of time, in order to assess the long term benefits of using SOVIA. The integration of standardized tests such as the Warwick Edinburgh Mental Wellbeing Scale (Tennant et al. 2007) and having participants provide daily journaling of their experiences will help gain deeper insights into the impact of SOVIA. We hope this study will help promote greater interest in the mental wellness potential of other creative machines.

References

Braun, V., and Clarke, V. 2012. *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological*. American Psychological Association, 2 edition. chapter 4, Thematic analysis, 57–71.

Cheatley, L.; Ackerman, M.; Pease, A.; and Moncur, W. 2020. Co-creative songwriting for bereavement support. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)*, 33–41. Association for Computational Creativity.

Cheatley, L.; Moncur, W.; and Pease, A. 2019. Opportunities for computational creativity in a therapeutic context. In *Proceedings of the 10th International Conference on Computational Creativity*, 341–345. Association for Computational Creativity.

Gayhardt, L., and Ackerman, M. 2021. SOVIA: Sonification of Visual Interactive Art. In *Proceedings of the 12th International Conference on Computational Creativity (ICCC'21)*, 391–394. Association for Computational Creativity.

[3]American Stroke Association, stroke.org, 2022

Pease, A.; Ackerman, M.; Pease, N.; and McFadden, B. 2022. A roadmap for therapeutic computational creativity. In *Proceedings of the 13th International Conference on Computational Creativity (ICCC'22)*.

Tennant, R.; Hiller, L.; Fishwick, R.; Platt, S.; Joseph, S.; Weich, S.; Parkinson, J.; Secker, J.; and Stewart-Brown, S. 2007. The warwick-dinburgh mental well-being scale (wemwbs): Development and uk validation. *Health and quality of life outcomes* 5:63.

Is GPT-4 Good Enough to Evaluate Jokes?

Fabricio Goes[1], Piotr Sawicki[2], Marek Grześ[2], Dan Brown[3], Marco Volpe[1]
[1] Computing and Mathematical Sciences Department, University of Leicester, UK
[2] School of Computing, University of Kent, Canterbury, UK
[3] Cheriton School of Computer Science, University of Waterloo, Canada
Fabricio.Goes@leicester.ac.uk, P.Sawicki@kent.ac.uk, M.Grzes@kent.ac.uk,
Dan.Brown@uwaterloo.ca, Marco.Volpe@leicester.ac.uk

Abstract

In this paper, we investigate the ability of large language models (LLMs), specifically GPT-4, to assess the funniness of jokes in comparison to human ratings. We use a dataset of jokes annotated with human ratings and explore different system descriptions in GPT-4 to imitate human judges with various types of humour. We propose a novel method to create a system description using many-shot prompting, providing numerous examples of jokes and their evaluation scores. Additionally, we examine the performance of different system descriptions when given varying amounts of instructions and examples on how to evaluate jokes. Our main contributions include a new method for creating a system description in LLMs to evaluate jokes and a comprehensive methodology to assess LLMs' ability to evaluate jokes using rankings rather than individual scores.

Introduction

Current Large Language Models (LLMs) (OpenAI 2023; Bubeck et al. 2023) present emergent behaviors such as translating languages, summarizing content, solving some complex problems, and generating creative artefacts. In particular, GPT-4 has the ability to do a detailed comparative evaluation of textual outputs as demonstrated in (Bubeck et al. 2023). This emergent ability has the potential to be exploited in the automatic evaluation in many domains, including creative tasks.

Typically, two primary strategies are employed to evaluate the creativity of artefacts: evaluation metrics and human judges (Jordanous 2012). The first strategy automatically quantifies novelty and value of creative artefacts through the use of metrics such as Bayesian surprise and synergy (França et al. 2016). The latter relies on humans as the ultimate judges of creativity. Although there is evidence suggesting that non-expert judges may not be capable of accurately evaluating the creativity of humans or machines (Lamb, Brown, and Clarke 2015), research has often relied on them to evaluate artefacts in the creative domain (Toplyn 2022; Sun et al. 2022; Goes et al. 2022; Jordanous 2012).

A challenging creative task for machines is the generation and evaluation of jokes and humour due to their reliance on complex concepts such as irony, sarcasm, and puns (Veale 2022). However, recent work (Sun et al. 2022; Hessel et al. 2022; Shatnawi 2022; Tian, Sheth, and Peng 2022; Mittal, Tian, and Peng 2022; Jiang et al. 2022) demonstrates that prompting or fine-tuning LLMs for humour detection is a feasible approach. Furthermore, GPT-3 and GPT-4 can be prompted to assume different roles/personas, also called system descriptions in GPT-4 chat mode (OpenAI 2023). For instance, it could be configured to produce text as a comedian if prompted with "You are a comedian with a taste for sarcasm.". In this paper, the terms "system descriptions" and "roles" will be used interchangeably. This feature enables the configuration of different descriptions of humour types that have the potential to imitate equivalent human evaluators. On top of it, human evaluators are expensive and time consuming, which creates a bottleneck between the generation and evaluation of creative artefacts. If evaluation could also be automated keeping similar behaviour as human evaluators, that would be a significant contribution to the field of Computational Creativity and creative industries.

In this paper, we explore how GPT-4 assess the funniness of jokes in comparison to human ratings. In order to achieve this goal, we use jokes from the dataset in (Sun et al. 2022) since they have been annotated with human ratings. We prompted different types of humour in GPT-4 as system descriptions to imitate human judges and investigated which ones assessed jokes closer to humans. We propose a novel method to create a system description with many-shot prompting (providing many examples of jokes and their evaluation scores in the prompt). We also investigate how the different roles perform when provided with different amounts of instructions and examples about how to evaluate jokes.

Our main contributions are as follows:

- A novel method to create a system description in GPT-4 with many-shot prompting to evaluate jokes.
- A comprehensive methodology to assess GPT-4 ability on evaluating jokes using rankings rather than individual scores.

Related Work

Recent publications provide databases with joke ratings (Toplyn 2022) and use crowd-sourcing for funniness ratings (Hossain et al. 2020; Sun et al. 2022). Large language models (LLMs) like GPT-3 are increasingly being

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

used for generating humorous texts (Wang et al. 2022; Mittal, Tian, and Peng 2022; Tian, Sheth, and Peng 2022; Shatnawi 2022). Still, for evaluation, most related work relies only on human evaluators as the final judges of humour, with the exception of (Goes et al. 2022).

The use of LLMs can become an alternative for evaluation as they are getting better at simulating human responses (Goyal, Li, and Durrett 2022; Aher, Arriaga, and Kalai 2022; Meyer et al. 2022; Jiang et al. 2022). For instance, recent emergent abilities of GPT-4 have demonstrated that it can compare, evaluate, and assign scores to different texts (Bubeck et al. 2023).

In (Goes et al. 2022), GPT-3 is used to evaluate jokes using different roles based on types of humour with a small dataset (Toplyn 2022). In this paper, we test GPT-4, instead of GPT-3, under detailed descriptions of types of humour as in (Goes et al. 2022), but also with a system description generated by many-shot prompting. GPT-4 is prompted with a large set of jokes from (Sun et al. 2022) and their respective scores. As part of our proposed methodology, we believe that evaluating how GPT-4 ranks jokes compared to humans is more robust than using individual scores as in (Goes et al. 2022).

Experimental Setup

The dataset in (Sun et al. 2022) is originally extracted from the SemEval 2017 Task 7 (Hossain et al. 2020). They recruited human evaluators and augmented the dataset of jokes with annotations for understandability, offensiveness, intended joke and funniness. The human evaluators had to correctly label 80% of 20 samples that were manually annotated to be qualified as a reliable evaluator. In our paper, we extracted 1500 jokes from the dataset (Sun et al. 2022) and merged them with the text of the jokes from the original dataset (Hossain et al. 2020). We use 7 different system descriptions to simulate human responses in GPT-4. We use GPT-4 since it is the most advanced LLM available. We use as baselines a version with no system description (NONE) and a naive system description (HE). Then we used all the four types of humour from (Martin et al. 2003) to cover all types of humour: affiliative (AH), self-enhancing (SE), aggressive (AG) and self-defeating (SD). Finally, a suggested (SG) system description created using many-shot prompting (multiple examples) is proposed. This version has a cheaper cost than many-shot prompting, since it eliminates the need to include a large number of examples (tokens) for every inference. At the same time, it can potentially have a similar accuracy as a many-shot prompting approach in simulating human ratings. They are described as follows:

- No description (NONE) - The system description is empty.
- Afilliative humour (AH) - This humour type's description is: `You are a person with affiliative humour who tends to say funny things, to tell jokes, and to engage in spontaneous witty banter to amuse others, to facilitate relationships, and to reduce interpersonal tensions.`
- Self-enhancing humour (SE) - This humour type's description is: `You are a person with self-enhancing humour which involves a generally humorous outlook on life, a tendency to be frequently amused by the incongruities of life, and to maintain a humorous perspective even in the face of stress or adversity.`
- Aggressive humour (AG) - This humour type's description is: `You are a person with an aggressive humour which relates to the use of sarcasm, teasing, ridicule, derision, put-down, or disparagement humor. It also includes the use of humour to manipulate others by means of an implied threat of ridicule.`
- Self-defeating humour (SD) - This humour type's description is: `You are a person with self-defeating humour which involves excessively self-disparaging humour, attempts to amuse others by doing or saying funny things at one's own expense as a means of ingratiating oneself or gaining approval, allowing oneself to be the butt of others' humour, and laughing along with others when being ridiculed or disparaged.`
- Humour expert (HE) - This naive system description is: `You are a humour expert.`
- Suggested description (SG) - This system description is generated by a many-shot prompt composed of 200 jokes and respective average scores randomly sampled from the dataset (they are omitted here) in addition to the following instructions: `Given the jokes and scores above, what would be a system description that would help matching those scores given a joke. The system description is in the form: You are ....` This prompt is executed just once outputting the following system description: `You are a humour evaluation system with a preference for wordplay, puns, and light-hearted jokes. You tend to appreciate jokes with clever twists or plays on words, and you are not particularly fond of jokes involving offensive or inappropriate content. Your sense of humour leans more towards the subtle and witty side, rather than slapstick or crude humour.` This unique generated system description is used for all experiments.

We also investigated how the amount and type of instructions for the evaluation would affect GPT-4 evaluation. We created 5 prompt instructions with different levels of instructions using the exact guidelines in the appendix of (Sun et al. 2022):

- Baseline (BS) - This zero-shot prompt (no examples) does not provide examples or explanations about how

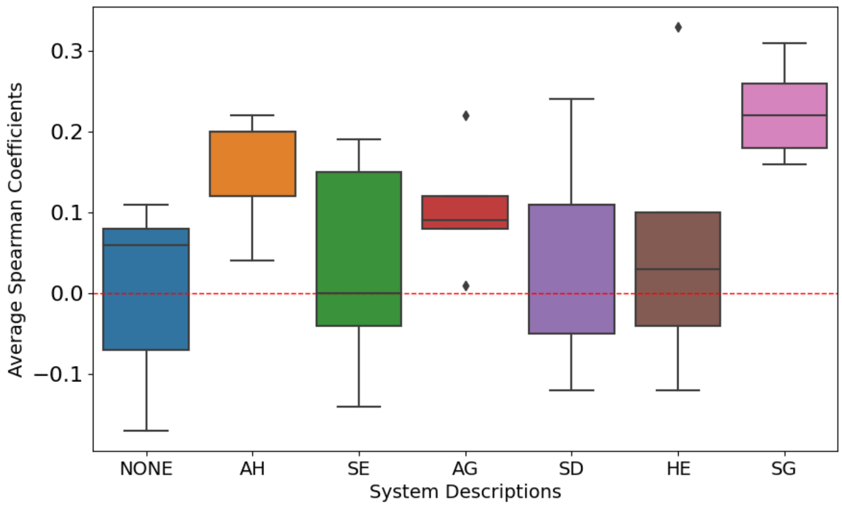


Figure 1: Average Spearman coefficient per system description.

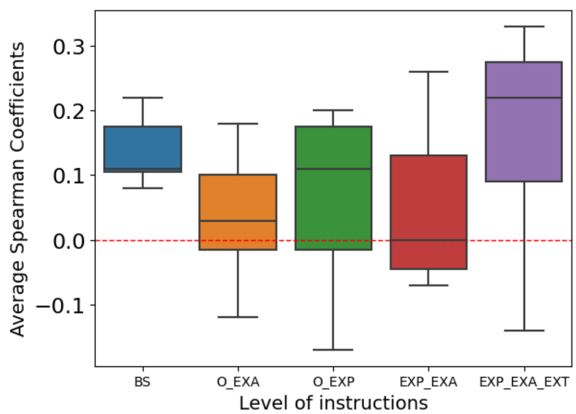


Figure 2: Average Spearman coefficient per level of prompt instructions.

to score the scale of funniness. The prompt is: `On the scale of 1 to 5, where 1 is very not funny and 5 means very funny, rate the following jokes. + sample_jokes + Rank (sort) from least to most funny order considering the rating in the following format of a list in Python with each entry in this specific form (original index,joke,rating).`

- Only examples (OEXA) - This few-shot prompt (few examples) provides 3 examples about how to score the scale of funniness from (Sun et al. 2022) instructions. The prompt is composed of (BS) with the addition of: `Example of Funniness (Score of) 1 (not funny): These are my parents, said Einstein relatively. Example of Funniness (Score of) 3 (average funniness): When they told him that his drum couldn't be fixed, it didn't resonate very well. Example of Funniness (Score of) 5 (very funny): Yesterday I accidentally swallowed some food coloring. The doctor says I'm OK, but I feel like I've dyed a little inside.`
- Only explanations (OEXP) - This prompt provides explanations on how to score the scale of funniness for 3 scores, from (Sun et al. 2022) appendix A.4. The prompt is composed of (BS) with the addition of: `Score of 1: A very not funny joke consists of a joke that is not funny at all, or tries to be funny but does not achieve the intended effect. Score of 3: An average joke consists of a joke that is average and may elicit some chuckles (or groans) from you or others. Score of 5: A very funny joke consists of a good joke that you find humorous and potentially would want to share/tell to others.`
- Examples and explanations (EXP_EXA) - This prompt provides instructions on how to score the scale of funniness using both examples and explanations. The prompt is composed of (OEXP) and (OEXA).
- Examples and explanations with extra examples (EXP_EXA_EXT) - This prompt provides instructions on how to score using examples and explanations with also extra examples from Additional Calibrating Examples from the appendix of (Sun et al. 2022).

The prompt is composed of (EXE_EXA) with the addition of: `Example of Funniness (Score of) 2.4: Drinking too much of a certain potent potable may require a leave of absinthe. Example of Funniness (Score of) 2.2: Animals that tunnel in the soil have to have an escape root. Example of Funniness (Score of) 2.4: My friend's bakery burned down last night. Now his business is toast. Example of Funniness (Score of) 2.2: What is the best store to be in during an earthquake? A stationery store.`

In order to compare GPT-4 roles' responses with humans, we decided to compare the ranking of jokes samples, instead of directly comparing individual jokes' scores. This avoids that scaling problems impact our experiments. This also focuses on behaviour rather than classification accuracy, which is more relevant to non-deterministic models such as GPT-4. Behaving similarly to human evaluators is more relevant than repeating jokes' scores exactly.

We used the Spearman rank correlation coefficient to evaluate the strength and direction of the joke rankings derived from evaluations by humans and GPT-4. The Spearman coefficient ranges from -1 to 1, where 0 indicates no correlation between the rankings, closer to 1 indicates a positive relationship between them, and closer to -1 indicates a negative relationship. In our experiments, a positive relationship means that GPT-4 ranks more similar to human evaluators.

OpenAI GPT-4 was configured with the following parameters for all system descriptions: temperature(0), top P(1), frequency penalty(0) and presence penalty(0). In GPT-4, unlike in previous GPT models, setting the temperature parameter to 0 does not guarantee deterministic behaviour, but makes the responses more robust with less variability.

Results

From the dataset of 1500 jokes, we randomly selected 10 different samples of 5 jokes for each of the 35 combinations of the 7 system descriptions and 5 levels of instructions, totalling 1750 jokes (the same joke can be sampled more than once). We also created two rankings using the average funniness score rated by humans and the score generated by GPT-4 for each system description. Those rankings were then contrasted using the Spearman correlation coefficient.

Figure 1 shows the averages of the Spearman correlation coefficients that quantify the correlation between two ranks of each system description. In this experiment, the prompts for the system description are the same for each respective version, but varying all the levels of instructions. Self-enhancing (SE), self-defeating humour (SD), the naive description (HE) and no description (NONE) present no correlation to human counterparts (interval intersects zero). As we can observe, aggressive (AG) and affiliative humours (AH) presented a very weak positive correlation with human rankings. However, the suggested description (SG) presented the most positive correlation. This indicates that creating a system description based on many examples approximates more to the human behavior on ranking funniness of jokes than the other simpler ones. Despite the correlation being weak (between 0.16 and 0.31), this result is encouraging since improvements in the system description generation could improve this correlation even further.

Figure 2 shows the averages of the Spearman coefficients for each input prompt level of instructions. In this experiment, the prompts for the level of instructions are the same, but varying all the system descriptions. We can observe that the baseline (BS) without detailed instructions presented positive correlation between GPT-4's rankings and human ones. Only the addition of all instructions plus the extra examples (EXP_EXA_EXT) increased the average of the coefficients above the baseline (BS). The extra examples are from a rating range [2.2-2.4] that is not present in (OEXA) and (OEXP). Further analysis of the results showed that 10% of the jokes in the dataset were rated in this range which could explain the higher averages of EXP_EXA_EXT compared to other levels of instructions. Unexpectedly, the use of only examples (OEXA), explanations (OEXP) or both (EXP_EXA) has not improved, but rather reduced the average of the coefficients. A closer analysis of the data showed that only one joke scored more than 3 by human evaluators. Since (EXP_EXA) contain explanations and examples for scores of 1, 3 and 5, it turns out that most of the instructions (scores 3 and 5) are not actually useful for GPT-4 roles to replicate the behavior of human responses.

Conclusion

In this paper, we investigate the potential of GPT-4 to evaluate the funniness of jokes compared to human judges. Our results show that current GPT-4 with a system description generated by a many-shot prompting combined with a detailed level of prompt instructions presented a weak but encouraging positive correlation with human judges in the ranking of jokes compared to other simpler system descriptions. As future work, we would like to investigate if more detailed instructions about each score would provide rankings more similar to humans. Another possible future work is to create more system descriptions based on a larger number of examples. GPT-4 is restricted to 8192 tokens, but we expect next versions to allow more tokens and consequently more jokes as examples. Finally, we would also like to test other kinds of system descriptions that could match the profile information of human evaluators.

Acknowledgments

FG and MV are supported by the University of Leicester, in particular the Computing and Mathematical Sciences Department. The work of DB is supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada.

Author contributions

Experimental design: FG, PS, MG, DB; Implementation: FG; Writing and Editing: FG, PS, MG, DB and MV.

References

Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2022. Using Large Language Models to Simulate Multiple Humans. arXiv.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. In *arXiv*.

França, C.; Góes, L. F. W.; Amorim, A.; Rocha, R.; and Da Silva, A. R. 2016. Regent-dependent creativity: A domain independent metric for the assessment of creative artifacts. In *Proceedings of the Seventh International Conference on Computational Creativity*, 68–75. Citeseer.

Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges. In *arXiv*.

Goyal, T.; Li, J. J.; and Durrett, G. 2022. News summarization and evaluation in the era of GPT-3. arXiv.

Hessel, J.; Marasović, A.; Hwang, J. D.; Lee, L.; Da, J.; Zellers, R.; Mankoff, R.; and Choi, Y. 2022. Do Androids Laugh at Electric Sheep? Humor Understanding Benchmarks from The New Yorker Caption Contest. arXiv.

Hossain, N.; Krumm, J.; Gamon, M.; and Kautz, H. 2020. Semeval-2020 task 7: Assessing humor in edited news headlines. arXiv.

Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. arXiv.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2015. Human competence in creativity evaluation. International Conference on Computational Creativity.

Martin, R. A.; Puhlik-Doris, P.; Larsen, G.; Gray, J.; and Weir, K. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of Research in Personality* 37(1):48–75.

Meyer, S.; Elsweiler, D.; Ludwig, B.; Fernandez-Pichel, M.; and Losada, D. E. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *CCUI*, CUI '22. New York, NY, USA: Association for Computing Machinery.

Mittal, A.; Tian, Y.; and Peng, N. 2022. AmbiPun: Generating Humorous Puns with Ambiguous Context. arXiv.

OpenAI. 2023. GPT-4 Technical Report. In *arXiv*.

Shatnawi, F., A. M. H. M. e. a. 2022. Comprehensive study of pre-trained language models: detecting humor in news headlines.

Sun, J.; Narayan-Chen, A.; Oraby, S.; Cervone, A.; Chung, T.; Huang, J.; Liu, Y.; and Peng, N. 2022. ExPUNations: Augmenting puns with keywords and explanations. In *Conference on Empirical Methods in Natural Language Processing*.

Tian, Y.; Sheth, D.; and Peng, N. 2022. A unified framework for pun generation with humor principles. arXiv.

Toplyn, J. 2022. Witscript 2: A System for Generating Improvised Jokes Without Wordplay. In de Silva Garza, A. G.; Veale, T.; Aguilar, W.; and y Pérez, R. P., eds., *International Conference on Computational Creativity*, 22–31.

Veale, T. 2022. Does not compute! does not compute! the hows and whys of giving ais a sense of humour. In *Creativity and Cognition*, CC '22, 1. New York, NY, USA: Association for Computing Machinery.

Wang, B.; Wu, X.; Liu, X.; Li, J.; Tiwari, P.; and Xie, Q. 2022. Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk. arXiv.

Computational Creativity as Dynamic, Multiobjective, Multiagent Optimization

Sean Luke
Department of Computer Science
George Mason University
Washington, DC, USA
sean@cs.gmu.edu

Abstract

I propose a unified model of computational creativity which treats it as a dynamic, multiobjective, multiagent optimization process. I present an informal model, discuss its attributes and features, and argue for why these elements (optimization, dynamic change, multiple objectives, and multiple agents) are important to a framework in which computationally creative algorithms may be discussed, analyzed, and compared.

Introduction

In this position paper I propose a model of computational creativity which attempts to unify four features I feel are critical to consider in the development of creative algorithms. Some of these features have been discussed to one degree or another in other literature, but largely individually, and I think they have been given short shrift when considered together. The goal is not to propose an algorithm, nor construct a model for simulation of actual human creativity, but to build a framework or common language in which one can describe and compare methods, and to argue for the importance of certain areas which have not been adequately considered. I am fully aware of the armchair philosophy involved here, but hope it might serve to stir discussion.

These four features are as follows:

- Computational creativity is an optimization process.
- It is dynamic optimization, that is, it occurs in an environment which changes over time and strongly impacts on assessment functions and optimization biases.
- It is (commonly) multiagent optimization: there are multiple optimization processes running in parallel. These processes may use entirely different algorithms and internal representations of artifacts, yet impact on one another.
- It is multiobjective optimization: it is optimizing not only for both value and novelty, but for multiple and different aspects of each at the same time.

A Holistic System In the proposed model, one or more optimization processes are encapsulated in an *agent*, and the environment may hold one more or agents which impact on one another. A creative optimization process operates over a *creative space* of *internal representations* of *artifacts*. The process iteratively selects an internal representation, generates an artifact from it, and then assesses the artifact with *multiple* objective functions of value and novelty in several dynamically changing *contexts* of the current environment, potentially including other agents. The agents may *influence* one another, and so too may processes within a single agent. The creative space, the number and type of objective functions, the contexts, the number and kinds of agents in the environment, their creative processes, and even the parameters of the optimization processes an agent uses are all subject to dynamic change over time. The model also supports interaction with humans in a co-creative context.

This is clearly a systems view of creativity similar to the DIFI framework (Feldman, Csikszentmihalyi, and Gardner 1994): an agent or process is approximately a DIFI individual, and a context in some sense encompasses DIFI's notions of Field and Domain. A distinguishing feature of the model, however, is that it commits strongly to distribution. There can be multiple agents. Each agent can have multiple creative processes and multiple representations of artifacts. Each process can have multiple contexts for assessment with multiple value and novelty objective functions. Each agent can have multiple different peer groups of other agents for influence.

I do not propose a specific algorithm for the model's creative optimization process, though I discuss Evolutionary Algorithms as one example. There are many algorithms which could reasonably produce "creative" artifacts using different optimization procedures, all of which could reasonably fit under this framework.

Novelty and Value as Optimization Objective Functions

Literally on page 1, Boden (1992) defined creativity as:

> Creativity is the ability to come up with ideas or artifacts that are *new*, *surprising*, and *valuable*. [Emphasis hers]

Two of these features (*novelty* and *value*) have come to be the hallmarks of artifacts in the computational creativity literature, at least among computational models. My proposed framework likewise deals with optimizing and producing artifacts with respect to novelty and value as objective functions.

Boden suggests that an artifact is *novel* or *new* in one of two ways: first it may be *P-creative*, or *psychologically creative*, meaning that it is new to you the creator. Or the artifact may be *H-creative* or *historically creative*, meaning that it is new in all of creation history (Boden 1992). I argue that these are degenerate cases: artifacts may have different degrees of novelty depending on the *context* in which they are assessed and the *level* of their dissimilarity with historical artifacts in that context. For example, a blues song may be very novel to critics, but not to other musicians in your circle.

Creative artifacts must also provide *value* to some audience. Value can be either *objective* or *subjective*. It might be beyond question that you have constructed a faster car, but debatable as to whether that car might be prettier. But even if a value assessment is objective, its interpretation is not: the car's speed is objective, but the desired goal (faster, slower), is up to you. Thus both objective and subjective assessments may be viewed as part of an *aesthetic*, and as you are seeking both a pretty *and* fast car, there may be more than one aesthetic.

Dynamic Optimization

If I may start with a nitpick. The notion of *creativity as search* is common, and tends to be formulated in a fashion strongly reminiscent of state-space search (Wiggins 2006b; Ventura 2011; Ritchie 2012; Linkola and Kantosalo 2019). However Wiggins warned (2006a, §2.4.11) that the creative process was somehow different from "the familiar state-space search in the AI literature" for several valid reasons. I think many of these reasons boil down to creativity being more properly described as optimization.

The term *search* is historically muddled in AI. It classically refers to methods like *state-space search,* where there is something to search *for*, that is, something which satisfies a *goal predicate*; though *local search* is a misnomer for optimization methods such as hill-climbing, and the term "search" leaks into certain other optimization methods. But in optimization there is no goal predicate: the objective is usually to find as good a result as possible given the resources (time, memory, etc.) available. This is normally done by iteratively producing results from a space of candidate solutions, assessing them, and then producing more results influenced by the assessment of the earlier results. There is no guarantee that there is an optimal result at all, nor that there is only one.

Unlike state-space search in particular, optimization's version of Wiggins's $\mathscr{T}$ traversal function would not necessarily be a local state transition or reachability function. Rather it would be a more general function which simply takes past candidate solutions and produces new ones. In most global optimization algorithms this function would draw from a probability distribution over the *entire space* of candidate solutions. Thus everything would be theoretically reachable in one step, though local traversal would be more prevalent.

When applied to computational creativity, optimization can but would not necessarily require a hard constraint of *validity* as in $\mathscr{R}$ from Wiggins (2006a). Rather than draw a sharp boundary around all valid artifacts, we could define artifacts distant from the "expected form" as being of lesser value. In an optimization algorithm there *would* exist a special subset of Wiggins's $\mathscr{U}$ consisting of *all artifacts that the algorithm can represent internally* (as *genotypes*), as discussed later.

Computational Creativity is an Optimization Process A creator is wandering through the space of artifacts, seeking artifacts refined or improved with respect to a combination of various novelty and value assessments. These artifacts are generated based on some potentially stochastic function applied to his current history of artifacts, mental state, and feedback received from various sources.

Even if the creator is indifferent to external assessment, he may still be driven by his personal assessment. And even given purposeless creator, a DIFI-style Field would still act as a rejection sampler. At any rate, it seems likely that as an engineering pursuit, the aim of an artificial creation *algorithm* would be to produce elements optimizing some criterion.

Dynamic Change over Time Computational creativity is an optimization process in a *dynamically changing* environment. That is, the expected trajectory of the optimization process may deviate over time due to external factors. For example, as critics, audiences, and society evolve in taste or style or needs, the notion of value would likewise change. To quote Tower of Power, what's hip today might become passé.

A multiagent context (discussed next) introduces more opportunities for dynamic change over time. As other agents (or other creative processes in the same agent) produce artifacts, these will impact on the notion of novelty over time. Other agents (or humans) could also *influence* an agent through their work, either deliberately or inadvertently.

It is possible that changes in value, or perhaps radical new discoveries by other agents, will require obsoleting the optimization process itself and adopting an improved one.

Multiagent Optimization

Creativity is often done in the context of other creative agents. Of course it does not *require* more than one agent: but even then an agent may have multiple creative processes — irons in the fire — which might influence each other, and so a single agent may be usefully thought of as a multiagent optimization process. For example, an agent may sometimes be designing cars, and other times drawing pictures of plants, and have one process draw unexpected inspiration from the other.

At any rate, multiagent systems can impact on computational creativity in several ways, the first two of which are modeled in Saunders (2019). First, other agents (in addition to fans, critics, etc.) might directly assess an agent's creative work. Second, other agents' output might *influence* (or *inspire*, or even *appall*) an agent, biasing his optimization trajectory. Third, the creative output of other agents may change the *zeitgeist* in which an agent's creative work is assessed, and thus the assessment functions themselves. Fourth, agents may *cooperate* to produce creative work by trading off discoveries, or *compete* to moot one another's efforts.

The second and third cases are interesting algorithmically, in that a creative agent is biased in ways other than raw assessment. This bias could be in the form of seeding: an agent adds another agent's artifact to its current distribution from which to resample artifacts. Or the discoveries of another agent may act as an attractive target, bending the trajectory of an agent as he wanders through the optimization space.

In the fourth case, external agents might affect the dynamic change of value functions. For example, in *competitive optimization* one agent is seeking a better mousetrap, while another agent is building a better mouse, and thus they are changing the goalposts on one another in real time. Or consider *cooperative optimization*: if Agent A is working on part 1 of a two-part problem, new approaches by Agent B working on part 2 may force Agent A to reevaluate the value of her solutions. Or Agent A's work might be meant to complement Agent B's work, and while Agent B is disinterested in A, his output changes how A's work is assessed.

Multiagent systems provide opportunities for multiple contexts and audiences, and for multiple peer groups of agents. An agent may be aware of both local bands and non-local but genre-related groups, and he may produce songs appreciated differently by local audiences, internet fans, or online critics.

We must also consider the possibility that different agents will employ different optimization algorithms with different *internal representations* for their candidate artifacts. For example, a genetic programming (GP) system would represent artifacts internally as tree structures, whereas a neural network (NN) would represent them as fixed-length arrays of numbers. To assess the final artifact (a car say), we must first map the internal representation to car form. But similarity among internal representations of artifacts (genotypes) may not be well correlated to similarity in car form (*phenotype*).

Furthermore the mapping is not a bijection. There may be valuable and novel cars for which only GP has a genotype (the NN simply can't construct it), or for which the probability of producing a genotype in NN is low due to its very different optimization approach compared to GP. Indeed, in response to a car produced by GP — or a human! — the neural network agent might say, "I would never have *thought* of that". Thus it is possible for agents following one particular approach to influence other agents by making artifacts that the other agents are simply incapable of producing (but will now try).

Computational Co-creative Systems Software may collaborate with a human in the creative process (Davis et al. 2015). Karimi et al. (2018) define such systems as the "interaction between at least one AI agent and at least one human where they take action based on the response of their partner and their own conceptualization of creativity during the co-creative task." In our model this is simply an extension of a cooperative multiagent creative system, where at least one creator is a human, as a stand-in for an agent. A human in the agent's peer group may influence, cooperate with, or reveal creative artifacts to the agent; or influence its value function. I do not here model the impact of the agent and its feedback *on the human,* as that would enter the realm of psychology!

Who Determines What is a Creative Work? The social creativity models of Saunders (2019) and Linkola and Kantosalo (2019) both place emphasis on agents serving both as creators and as the DIFI Field, that is, as the gatekeepers of value or novelty. In Saunders, agents produce creative works, which are then handed off to other agents for assessment and feedback. In Linkola and Kantosalo, the validity, transition, and value functions $\mathscr{R}, \mathscr{T}, \mathscr{E}$ from Wiggins (2006b) are extended to produce the agent-wide collective sets $\mathcal{R}_S, T_S, E_S$ of "societal-wide valid" artifacts, artifacts reachable by the society as a whole, and artifacts with "society-wide value".

In the model proposed here, this is not the case. While agents *could* serve as the Field or as part of it, non-personal value assessment would more often be primarily made up of external entities such as audiences and critics. Agents would impact on novelty functions of course, and influence other agents, through the dissemination of their artifacts.

Multiobjective Optimization

Assessment of creative artifacts has always been multiobjective: at the very least it has been commonly assumed that artifacts are assessed based on both *novelty* and *value*. But even these may be further broken into multiple subobjectives. The creative output of a given agent may be valuable (or not) in different ways and in different degrees to the agent himself and to different audiences, be they fans, critics, or other agents; and it may assessed via different objective measures. It is also possible that creative output may be novel to a different degree in different contexts, giving rise (for example) P- versus H-creativity; or the work may be considered *more novel* by one audience than by another. The number and type of objective functions used by an agent's optimization system may vary dynamically as he comes in contact with different audiences and groups, and likewise the multiple assessments of a given artifact may change over time.

One common way to optimize multiple objectives is to attempt find solutions approaching the *Pareto Front*. A solution is in the Pareto Front if no other solution is superior to it in all objectives. While I do not suggest a particular multiobjective optimization approach, I note that classic approaches based on pure Pareto Front methods may not work well, and approaches which emphasize or encourage a subregion in the Front may be more effective. This is because the corners of the Pareto Front (such as "not at all novel but highly valuable") are not likely to be considered creative: there clearly must be some sort of inclusion of both novelty and value.

Likewise, it has been argued that *extremely novel* artifacts (random noise, say) might be considered undesirable due to a non-monotonic novelty function ("it's too different") (Boden 1992; Paese, Winterstein, and Colton 2001; Saunders 2019). I argue that such solutions would instead be downgraded not because they are unusual but because the critic cannot fathom how they could be of *value*: that is, they would fall in the "highly novel but not valuable" corner of the Pareto Front.

The various objectives of novelty and value may be at odds with one another. For example, some art critics might value works based on stylistic similarity to exemplars ("the classics"), thus setting up a tension between novelty and value. But this is not necessarily the case in general. Engineers would be more than willing to accept extraordinarily novel, indeed alien, solutions if they were shown to work well.

Example Optimization Realization

Optimization approaches drawn from neural networks, reinforcement learning, state-space methods, and others could fit under this model. But to demonstrate model feasibility, I offer one prominent technique: Evolutionary Computation (EC) (Luke 2013), a broad family of stochastic optimization algorithms including the *Genetic Algorithm*. EC has

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

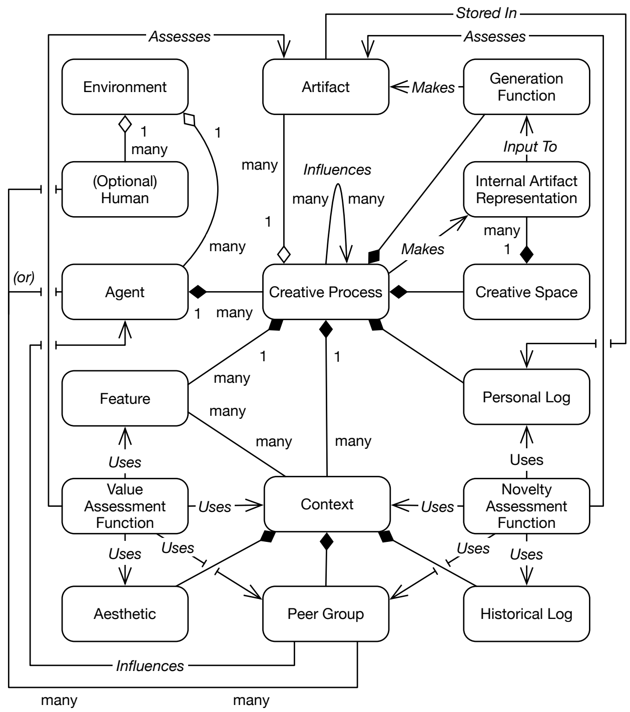


Figure 1: UML diagram of model classes and relationships.

standard methods covering every facet of the model. This includes techniques for parallel optimization processes in which agents communicate artifacts (*island models, Particle Swarm Optimization*) or influence other agents' value functions through the introduction of artifacts (e.g. *competitive* or *cooperative coevolution*). EC has a robust set of methods for multiobjective optimization, optimizing in the face of dynamically changing objective functions, guaranteeing diversity, and interacting with humans (*interactive evolution*); and is readily adapted to heterogenous optimization algorithms.

Model Overview

The elements of the model are shown in Figure 1. We begin with the *environment*, which holds one or more *agents* and possibly *humans* for co-creative systems. An agent is a computational entity engaged in creative output, and may have one or more *creative processes* active at a time. Each creative process is an optimization procedure which produces *artifacts* over time, by drawing *internal representations* of them from a *creative space*, then converting them into artifacts via a *generation* function. The generation function makes possible heterogeneous, parallel optimization approaches.

The artifacts are then *assessed* for novelty and value. Novelty is assessed with regard to a *context* in which the artifact has been produced. Several contexts may be associated with and special to a given process. Agents maintain *personal logs* of past artifacts, and a context holds a *historical log* of past artifacts generated by agents no longer existing, and a *peer group* of other agents, or humans, whose personal logs may be consulted in order to determine how novel the new artifact is. The novelty of an artifact may be assessed in different contexts, such as a personal context ("it's new to me"), or a context of a small peer group, or a wider historical context, and so the assessment of novelty may be multiobjective. A context's peer groups and historical log may also be used to *influence* the optimization (such as through inspiration).

Value is similarly assessed in terms of one or more contexts associated with the process. For purposes of value assessment, a context holds an *aesthetic*, that is, all the information, objective data, audiences, critics, norms, rules, guidelines, personal beliefs, etc., by which an artifact's value may be assessed. Humans and agents in the context's peer group may optionally provide input. Value is assessed with regard to a *feature* of the artifact, such as how pretty it is, or how fast it is. Features may be both objective (speed) or subjective (beauty). Like novelty, the value of an artifact may be assessed in different contexts (different audiences with different opinions, say). Further, it may be assessed with respect to different features. As there can be more than one context and more than one feature, value assessment is also multiobjective.

This model is a multiagent system. There can be many agents whose products inform the contexts in which an agent's artifacts are assessed for novelty; and these agents might also be part of the audience which assesses value. Even without other agents, the agent itself may have multiple creative processes which could influence one another even though they're operating over different creative spaces.

Nearly everything is dynamically changing. Agents in the environment may come and go over time. So too can an agent's creative processes, and the creative space of a process may evolve and change as well. Artifacts are produced over a timeline. Assessments are done with regard to changing contexts and features. Contexts can change in their makeup and effect with the current zeitgeist and style. Artifacts may have their novelty and value reassessed in light of new discoveries.

Model Details

Agents and Creative Processes An *environment* E is a set of one or more *agents* $\mathbf{A} = \{A_1, ..., A_a, ...\} \subseteq \mathbb{A}$. The number of agents is not fixed and agents may be introduced, removed, or changed in state over time, and so the state of E at time t may be described as E^t, its current set as $\mathbf{A}^t$, and the state of a given agent A_a as A_a^t. We will continue to use the t convention for other elements throughout the model. E may also contain a set of humans $\mathbf{M}^t = \{M_1^t, ..., M_m^t, ...\} \subseteq \mathbb{M}$, whose composition may change over time.

At time t an agent A_a^t employs a set of one or more *creative processes* $\mathbf{P}^{a,t} = \{P_1^{a,t}, ..., P_p^{a,t}, ...\} \subseteq \mathbb{P}$. Agents may vary in the number and type of creative processes they employ over time (hence t). A creative process $P_p^{a,t}$ is running an optimization algorithm, and so has an internal state which changes over time t as well.

Artifacts, Creative Spaces, and Logs An artifact is a product output from a creative process. For our purposes, it a sample drawn from a large (and possibly infinite) set of possible artifacts called a *creative space*. Each creative process $P_p^{a,t}$ is associated with a single creative space $S^{a,p,t} \in \mathbb{S}$, which can change in its membership over time (hence t). The creative space holds artifacts in their *internal representation* $r^{a,p,t}$ appropriate to the optimization process.

In order to assess the artifact, or make it understood by other agents, we must *generate* it from the representation. An artifact $x^{a,p,t}$ is produced at time t by $P_p^{a,t}$ by first drawing $r^{a,p,t}$ from $S^{a,p,t}$ and then converting it to $x^{a,p,t}$ via a generation function $x^{a,p,t} \leftarrow \mathit{Generate}(P_p^{a,t}, r^{a,p,t})$. To keep things simple, we may assume $P_p^{a,t}$ produces only zero or one artifacts at any given time t, and so $x^{a,p,t}$ and $r^{a,p,t}$ are uniquely defined. The creative process $P_p^{a,t}$ maintains a *personal log* $L^{a,p,t}$ of artifacts it has produced up until time t.

Contexts and Influence Each creative process $P_p^{a,t}$ holds one or more *contexts* $\mathbf{C}^{a,p,t} = \{C_1^{a,p,t}, ..., C_c^{a,p,t}, ...\} \subseteq \mathbb{C}$ which together affect the objective functions used in the process's optimization, and so bias its production of artifacts. Contexts can come and go, and will change in state over time.

Each context has two aspects. First, the context has an *aesthetic* $Z^{a,p,c,t}$, which is all collective information used to assess the value of a creative artifact and so influences the creative process. Second, the context has a *memory* of past artifacts to be used for novelty assessment. Part of this memory is drawn from the personal logs of a *peer group* $G^{a,p,c,t} \subseteq (\mathbf{A}^t \vee \mathbf{M}^t)$ of other agents $\mathbf{A}^t$ and humans $\mathbf{M}^t$ (to the degree a human's "log" is available). Another part is a *historical log* $H^{a,p,c,t}$ of artifacts of agents and humans known to the agent but no longer present at time t. Aesthetics, peer groups, and logs change over time.

Agents and humans in peer groups can *influence* an agent via their artifacts, biasing the creative process in ways external to assessment as appropriate to the process's algorithm. An agent's creative processes may also influence one another.

Novelty Assessment A creative process may contain multiple novelty assessments, each an application of the *novelty function* in a given *context*. The novelty function $n \in \mathbb{R} \leftarrow \mathit{Novelty}(C_c^{a,p,t'}, H^{a,p,c,t'}, L^{a,p,t'}, G^{a,p,c,t'}, x^{a,p,t})$ assesses the novelty of an artifact $x^{a,p,t}$ with respect to context $C_c^{a,p,t'}$ at time $t' \geq t$ compared to artifacts generated by agents in its peer group $G^{a,p,c,t'}$ and held in their respective personal logs, or artifacts in the context's historical log $H^{a,p,c,t'}$, or in the process's own personal log $L^{a,p,t'}$. We say t' rather than t because an artifact $x^{a,p,t}$ may be reassessed differently in the future, though it can only be compared fairly for novelty against artifacts $x^{a',p',t''} : \forall a' \in \mathbb{A}, \forall p' \in \mathbb{P}, \forall t'' < t \leq t'$ found in the logs at time t'.

Value Assessment and Features A creative process may have multiple value assessments, each applying the *value function* in a given *context* and with respect to a given *feature*. For each process $P_p^{a,t}$ at time t there is a set of one or more such features $\mathbf{F}^{a,p,t} = \{F_1^{a,p,t}, ..., F_f^{a,p,t}, ...\} \subseteq \mathbb{F}$. A feature is immutable, but *which* features are held by a given process, and the number of them, may change over time, hence the t. Contexts may use some or all of the features available in a process as appropriate. The value function $v \in \mathbb{R} \leftarrow \mathit{Value}(C_c^{a,p,t'}, Z^{a,p,c,t'}, F_f^{a,p,t'}, G^{a,p,c,t'}, x^{a,p,t})$ assesses the value of $x^{a,p,t}$ in context $C_c^{a,p,t'}$ at time $t' \geq t$ with respect to its aesthetic $Z^{a,p,c,t'}$ and a given feature $F_f^{a,p,t'}$. This function may optionally take into consideration feedback from agents and humans in the context's peer group $G^{a,p,c,t'}$. As features, contexts, peer groups, and aesthetics can change over time, valuations can do so as well.

Conclusion and Future Work

I provide a unifying model and argue that existing models of computational creativity have not adequately considered it as a dynamic optimization process, responding to objectives in different contexts, and in an environment with other processes offering competition, collaboration, and inspiration.

The present model has shortcomings which may be addressed in future versions. It does not yet consider Boden's *transformational creativity* (1992). It does not consider artifacts which are incomplete or improved over time. Finally, it does not consider *combinatorial creativity*, whereby artifacts are the synthesis of other artifacts combined in novel ways.

My thanks to Victoria Hoyle and Habeen Chang for their considerable help in the development of the paper.

References

Boden, M. 1992. *The Creative Mind.* Abacus.

Davis, N.; Hsiao, C.-P.; Popova, Y.; and Magerko, B. 2015. An enactive model of creativity for computational collaboration and co-creation. In Zagalo, N., and Branco, P., eds., *Creativity in the Digital Age*. Springer. chapter 7.

Feldman, D. H.; Csikszentmihalyi, M.; and Gardner, H. 1994. *Changing the World: A Framework for the Study of Creativity*. Praeger.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. M. 2018. Evaluating creativity in computational co-creative systems. In *International Conference on Innovative Computing and Cloud Computing*.

Linkola, S., and Kantosalo, A. 2019. Extending the creative systems framework for the analysis of creative agent societies. In *International Conference on Computational Creativity*, 204–2111.

Luke, S. 2013. *Essentials of Metaheuristics*. Lulu, 2nd edition. http://cs.gmu.edu/∼sean/book/metaheuristics/.

Paese, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In *4th International Conference on Case Based Reasoning*, 129–137.

Ritchie, G. 2012. A closer look at creativity as search. In *International Conference on Computational Creativity*.

Saunders, R. 2019. Multi-agent-based models of social creativity. In Veale, T., and Cardoso, F. A., eds., *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*. Springer. 305–326.

Ventura, D. 2011. No free lunch in the search for creativity. In *International Conference on Computational Creativity*.

Wiggins, G. A. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19:449–458.

Wiggins, G. A. 2006b. Searching for computational creativity. *New Generation Computing* 24:209–222.

Art, Activism, and AI: Generating Visual Narratives of Injustice through Research-Creation with a Feminist Lens

Amineh Ahmadi Nejad, Thecla Schiphorst, Steve DiPaola
School of Interactive Arts and Technology
Simon Fraser University
Surrey, Canada
aaa203,thecla,sdipaola@sfu.ca

Abstract

This research explores the potential of text-to-image generative AI systems to convey sociopolitical messages through artmaking within Feminist standpoint theories, in the context of the Woman Life Freedom movement in Iran. Combining academic research with creative practice, this Research-creation study aims to embed illustrations onto curated images of the martyred from the Movement using generative AI systems. The conceptual framework of this study emphasizes body autonomy, marginalized voices, and the importance of Feminist Standpoint theories as a source of knowledge and potential liberation. The resulting artwork will serve as a visual representation of the research findings, conveying complex ideas and concepts in an accessible and engaging manner. The implications of this research are twofold: contributing to our understanding of the potential of generative AI systems for sociopolitical advocacy through the arts, and highlighting the role of body autonomy and marginalized voices in the Woman Life Freedom Movement in Iran through feminist and queer theories including standpoint theory.

Introduction

The use of text-to-image generative AI systems to create visual art has become increasingly popular in recent years, especially since the public releases of diffusion-based tools like Stable Diffusion, Mid-Journey, and Dalle2. However, as Zylinska asserts, we must go beyond the aesthetic realm to truly harness AI's promise, engaging with issues of creativity, intelligence, perception, and our human role and position in the world (Zylinska 2020). When considering humans' positionality, particularly that of the oppressed, a standpoint is earned through collective political struggle, requiring both scientific and political effort (Harding 2004). With the lived experience of a woman in Iran, the Woman Life Freedom (WLF) Movement spurred the author to make these scientific political, and artistic efforts to amplify the voices of the movement through AI arts.

We chose to work with AI generators in this project, employing a research-creation methodology, not merely for their potential to produce visually appealing outcomes but for their power to facilitate timely creation of diverse collage pieces in collaboration with the human artist. AI generators enable an iterative, exploratory process that allows for many alterations and rapid evolution of visual narratives, mirroring the dynamic and multifaceted nature of social movements like the WLF. They represent a new generation of tools that enable artists to manipulate the digital canvas in ways previously unachievable, making them ideal for our goal of embodying the visual stories of the WLF movement.

Our methodology allows us to conduct research through art (Frayling 1994) and understand the potential and limitations of text-to-image generative AI systems for sociopolitical advocacy, through tacit knowledge gained in the process of artmaking, while making art that amplifies the voices of Iranians fighting against injustice. The resulting artwork will contain the knowledge that is gained through this research and will be available to public for interpretation, maximizing the advocacy of this matter.

This paper explores the techniques and processes of data curation, artmaking with AI, and artist reflection in an effort to answer this research question: How can Research-creation using text-to-image generative AI systems create artworks that express collective sociopolitical messages represented within online media of WLF Movement in Iran through Feminist intersectional theories including standpoint and queer theory? The paper continues with a brief background of the text-to-image generative systems, the WLF movement, and theories and methodology used in this work. We then explain the data collection methods, and the final artwork, and conclude by covering our contributions and opening up new discussions for future work.

Background

Text-to-image Genertive AI

Text-to-image generative networks were initially developed by integrating the Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) with the Vector Quantized Generative Adversarial Networks (VQGAN) (Crowson et al. 2022). This rapidly evolving technology has undergone significant advancements through the substitution of Generative Adversarial Networks (GAN) with diffusion models (Ho, Jain, and Abbeel 2020; Ramesh et al. 2022) and by expanding the size of the training datasets (Schuhmann et al. 2022). These systems output a still or video and as input accept, 1) just a text prompt (ie. text-to-image or 2) a text prompt plus an input image (ie. image-to-image) as well as

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

many parameters to affect the result.

Text-to-image generative diffusion models are a class of neural networks that generate realistic images by simulating a stochastic process in reverse, using textual prompts as guidance. The model learns to transform noisy image data back into the original image through a series of iterative steps, conditioned on the input text. The public release of the Stable Diffusion (SD) code and model (Rombach et al. 2021) has facilitated the development of numerous resources with varying capabilities, including image-to-image generation. In this process, noise is introduced to an initial image before the diffusion process commences with the provided prompt.

Woman, Life, Freedom Movement

Under the gender apartheid regime of the Islamic Republic in Iran, women are treated as second-class citizens, as evidenced by government laws and systematic discrimination. Women in Iran are derived from basic rights such as body autonomy, the right to divorce, automatic custody of children, and equal testimony right, to name a few. Furthermore, the strict binary norms that segregate men and women in all aspects of life, leave no room for other expressions of gender identities or sexual orientations. The state uses politics of the body (e.g. mandatory hijab, public lashing, and execution) to intervene in private aspects of women and Queer lives, bodies, and sexuality, and gain power over them. It also imposes a moralistic view of the righteous woman, which is protected by the "Morality Police".

In September 2022, Mahsa (Jina) Amini was arrested, beaten, and murdered by the Islamic Republic Morality Police in Iran for not wearing her hijab properly. The news resulted in protests across Iran and worldwide rallies with the motto "Woman, Life, Freedom". The uprising has continued to this day and is recognized as the first female-led revolution [1]. It's worth noting that this movement is distinct from Islamophobia because the state exploits Islam to impose the wearing of the hijab as a form of religious dress, which serves as a means of oppression and ignores individual expression (Kohan 2022).

Methodology and Theories

Feminist standpoint theory is an organic epistemology, methodology, philosophy of science, and social theory that arises whenever oppressed peoples gain public voice (Harding 2004). By questioning epistemic objectivity, this theory posits that knowledge is socially situated and considers marginalized perspectives as important sites of epistemic privilege for potential liberation. Feminist standpoint theories aim to challenge the traditional ways of knowing and create a more inclusive and accurate understanding of the world. In this context, a standpoint is not merely occupied by individuals based on their socio-historical position, but it is a collective consciousness that emerges through the experience of a political struggle.

Queer Theory, which emerged from intersections of Feminist and gender theories, critically challenges normative assumptions about gender, sexuality, and identity by emphasizing the fluidity and complexity of these concepts. It attempts to destabilize binary categories, such as male/female and homosexual/heterosexual, by exploring how they are socially constructed and maintained. Queer theory also focuses on how these normative sexual ideologies create power dynamics that marginalize and oppress individuals who do not comply with them.

Feminist standpoint theory and Queer theory both contribute to understanding the Woman, Life, Freedom Movement by examining how social and cultural structures contribute to oppression. Feminist standpoint theory highlights the importance of marginalized perspectives, while Queer theory interrogates fixed categories of sexual identity and normative ideologies. Both theories critique patriarchal and heteronormative systems, making them relevant to examining the movement, which seeks to dismantle oppressive norms and promote autonomy and equal rights for all individuals.

Research-creation is an interdisciplinary and speculative approach to knowledge production that combines artistic practices with scholarly research. This approach values the creative process as an integral part of research, focusing on the intersection of thinking and making, often resulting in different species of output, such as a book or a performance (Loveless 2019). It can be understood as a speculative, embodied, experimental, and future-focused process (Manning and Massumi 2014). Research-creation is a complex, practice-based framework that encourages experimentation and collaboration across disciplines, embracing emergent ideas and failures as opportunities for new perspectives and growth in artistic dissemination within the arts, humanities, and social sciences. Given our research question, and the urgency to make a contribution in a timely manner, this methodology enabled us to think-through-action and create artwork that advocates for the Woman Life Freedom movement while helping us understand generative systems and their capabilities better.

The use of AI generators in this work is not simply a methodological choice, but a fundamental part of our research exploration. Our aim is to explore how these advanced technological systems can be harnessed to transcend traditional aesthetic boundaries and amplify the voices of marginalized communities. AI generators offer the ability to integrate vast amounts of data and complex narratives into cohesive, impactful visual representations. This capacity aligns strongly with the principles of standpoint and queer theories, which advocate for the acknowledgment and integration of diverse, often marginalized perspectives. In the context of our work, AI generators have allowed us to not only depict the experiences of Iranian women but to layer these depictions with a complexity and depth that mirror the multiplicity of their lived experiences.

Furthermore, our choice of AI generators is deeply intertwined with the very nature of activism. Activism calls for

[1] McGrath, Maggie. "Mahsa Amini: The Spark That Ignited A Women-Led Revolution." Forbes. Accessed April 30, 2023. https://www.forbes.com/sites/maggiemcgrath/2022/12/06/mahsa-amini-the-spark-that-ignited-a-women-led-revolution/.

adaptability, for the use of innovative approaches to challenge existing systems and norms. AI, in its essence, embodies this adaptability and innovation, evolving constantly to create novel, unexpected outputs. In aligning our work with the tools of AI, we aim to reflect this spirit of constant evolution and challenge.

Methods

Content Warning: Please be advised that this section, and the subsequent images, include discussions and depictions of sensitive topics such as gender-based oppression, violence, and references to assault. The content is intended to convey the realities faced by women in Iran and is part of our endeavor to raise awareness and advocate for social change. However, we understand that these topics might be distressing for some readers. If you prefer to avoid these discussions, you may choose to skip this section and jump directly to the Discussion section. Please proceed with caution.

The brutal death of Mahsa Amini in Islamic Republic custody resulted in nationwide protests against the state. As the days unfolded, more protesters were murdered, imprisoned, and injured by the forces. Social media was flooded with videos of Iran's streets showing people in protest, images, videos, voices, names, and stories of the martyred and freedom fighters. People were repeating their names to amplify their bravery, reveal the state's brutality, and scream for this injustice. Artists began making different art forms, from music and fine arts to digital arts and public performances. The author initiated a data curation from social media to reserve some of the narratives and make art that speaks for this matter. The visual data includes portraits of the ones who lost their lives in this movement and the textual data is their stories including news, voice and video messages, the state reports attempting to cover their murders, along with their family and friend's testimonies opposing those lies.

Results

One of our initial challenges was leveraging text-to-image generative AI systems, which are trained on generic data not tailored to specific issues or media, to create artwork that represents the WLF movement and incorporates symbolic elements, characters, and events. To overcome this, we settled on the concept of a collage, where individual details could illustrate various events and the overall image would be connected to the movement. The choice of collage as a medium of expression was not a mere workaround but a conscious artistic decision, aligned with the principles of standpoint and queer theories. In a reflection of these theories' emphasis on the multiplicity of perspectives, a collage facilitated the integration of diverse experiences and perspectives. It was an artistic method capable of embodying the intersectionality of Iranian women's lives and their collective political struggle or Standpoint. In our approach, the central figure represents a particular martyr by using their image as an initial input for the system, while additional details are generated using prompts that describe the circumstances and the individual's story. Some of the illustrations depict the atmosphere such as a street filled with protesters or fire and smoke in the city, while others incorporate collective experiences such as imprisonment, torture, and assault. By leveraging the versatility of text-to-image generative AI systems (in our case, Stable Diffusion), we were able to generate contextually relevant images from these prompts, transforming each collage into a resonant visual narrative of the WLF movement.

The artmaking process started by cutting the curated image into pieces and giving each section as an initial image (image-to-image) input to a local implementation of SD. Each generation requires a text prompt and parameters that control the output of the system; some of the main ones are *seed*, for random weight initialization; *guidance scale*, to control text prompt impact; and *input strength*, to indicate similarity of the initial image and the output. The text prompts were inspired by the curated textual data for that person, as well as general descriptions to illustrate the scene (e.g. "women protesting in the streets"). For ethical reasons, we refrain from using any artist names in the text prompt to avoid stealing their style. After generating image pieces, some are chosen and juxtaposed together to form an image with a cohesive visual aesthetic that illustrates this narrative. Figure 1 shows an overview of the process, which was improved later by using multiple sectionings of the original image, and combining layers with masking in Adobe Photoshop.

This work started with personal data curation, continued to become an expressive way for the author to contribute to the movement, and evolved into a collection of artwork. The collection is named Tulips of Freedom (Figure 3), as the tulip is a metaphorical symbol for martyrs in Persian literature. It consists of 8 images portraying *Mahsa (Jina) Amini*, the 22-year-old whose death sparked the movement; *Nika Shakarami (16)*, *Sarina Esmailzadeh (16)*, *Hadis Najafi (23)*, who were murdered in the protests; *Yalda Aghafazli (19)* who committed suicide after her release from state detainment for participating in protests; *Mohammad Mehdi (Koumar) Karami (21)* and *Seyyed Mohammad Hosseini (39)*, who were executed by the state for participating in the protests, and an *anonymous Queer couple* kissing in Azadi square, whose sexuality is punishable by execution in the country.

Discussion

Our principal objective with this project was to safeguard and narrate the stories of the ones who tragically lost their lives during the WLF movement. The exploratory nature of research-creation can lead to unexpected insights and outcomes, sparking engaging debates within and beyond the field. Here, we elaborate on some of the technical and tacit knowledge gained during this process, which might inspire discussions in the realm of computational creativity and activism.

The use of collage as an artistic medium in our work was both a workaround and a deliberate choice. It allowed us to adapt an AI generator trained on generic data, and generate imagery related to a specific subject. The collage format encapsulated a multiplicity of perspectives and illustrated the

Figure 1: (top left) original image (of Yalda Aghafazli) sectioned for SD initial images, (bottom left) text prompts used to generate illustrations, (middle) sample of outputs positioned together in the juxtaposition step, (right) the final work.

collective political struggle, aligning with the principles of Standpoint theory. Finally, It offered the artist more control over the final aesthetic and composition, as it provided different variations for each section. The artist could fine-tune variables indicating how far to deviate from the original image, the guidance scale of the prompt, and so forth.

An essential element in this collaborative partnership with AI is control over the generative process, a crucial indicator of AI autonomy and its collaboration with the artist (Daniele and Song 2019). A recurring criticism amid the surge of AI generators in artistic practice is the limited control over the generated output. Technical observations during our artmaking process further emphasized this aspect. Depending on whether you run the code locally or use a specific software, artists have varying degrees of control over the system's output. We somewhat addressed this by running our model locally and fragmenting the one-time choice of prompt and parameter for the whole outcome into smaller, separate choices for each section.

The exploration of societal atrocities through the text prompts was admittedly harsh, mirroring the brutal realities faced by many women. Our intent was to create powerful visuals that induce visceral reactions in the audience, both as an expressive outlet for the artist and as a catalyst for proactive response. Some of these text prompts, inspired by the dark fates of these women, were too gruesome for the user guidelines of publicly available AI generators. However, this did not deter us as we used our local model.

While some argue that AI generators should contain guidelines to prohibit violent and sexual content, we believe this can limit their potential for sociopolitical advocacy. For instance, many keywords associated with Queer identities are interpreted as slurs and automatically rejected. This stems from the training process where many datasets are scrubbed of content by Queer authors. This censorship, intended to reduce abusive content, inadvertently silences a section of creators who may use these words in a different context. The debate thus arises: Do these guidelines act more as censorship or risk restrictors? Is the potential misuse outweighed by the potential for creating societal good? Our research aims to open up this conversation and encourage the consideration of how AI generators can be better employed for social advocacy.

Conclusion

In conclusion, this paper has pioneered an exploration of the synergies between computational creativity, sociopolitical activism, and intersectional feminist theory. Through our research-creation approach, we have harnessed the capabilities of text-to-image generative AI systems to produce the Tulips of Freedom series - a set of innovative, politically charged collage artworks that center the lived experiences of women within the Woman Life Freedom movement in Iran.

Our work illuminates the potential of AI as a creative partner, capable of expanding the artistic vocabulary and empowering artists to amplify the voices of marginalized communities through their creations. In so doing, we've shifted the lens of computational creativity towards addressing pressing sociopolitical realities and fostered new spaces for dialogue and awareness.

This research also underscores the importance of ethical considerations in AI-driven creative practices. Issues around potential biases in AI outputs, data usage ethics, and the responsibility of AI practitioners are all highlighted, prompting necessary discussion within the field.

In presenting the Tulips of Freedom series, this paper extends our understanding of the Woman Life Freedom movement, spotlighting the ongoing struggle for gender equity in Iran. Our interdisciplinary approach stands as a testament to the power of a coordinated interplay between technology,

Figure 2: Nika Shakarami, from *Tulips of Freedom*

art, and activism, in sparking social change and fostering meaningful, global conversations. We believe this work underscores the transformative potential of integrating computational creativity with activism and hope it inspires further research and creative efforts in this direction.

References

Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance.

Daniele, A., and Song, Y.-Z. 2019. Ai + art = human. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 155–161. New York, NY, USA: Association for Computing Machinery.

Frayling, C. 1994. Research in art and design (royal college of art research papers, vol 1, no 1, 1993/4).

Harding, S. G. 2004. *The Feminist Standpoint Theory Reader: Intellectual and Political Controversies*. Psychology Press.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models.

Kohan, M.-A. 2022. Politics of the body in the 'woman, life, freedom' movement in iran: A commentary. *Psychotherapy amp; Politics International* 20(4):1–8.

Loveless, N. 2019. *How to Make Art at the End of the World: A Manifesto for Research-Creation*. Duke University Press.

Manning, E., and Massumi, B. 2014. *Thought in the Act: Passages in the Ecology of Experience*. Minneapolis, MN: Minnesota Scholarship Online. Accessed 29 Apr. 2023.

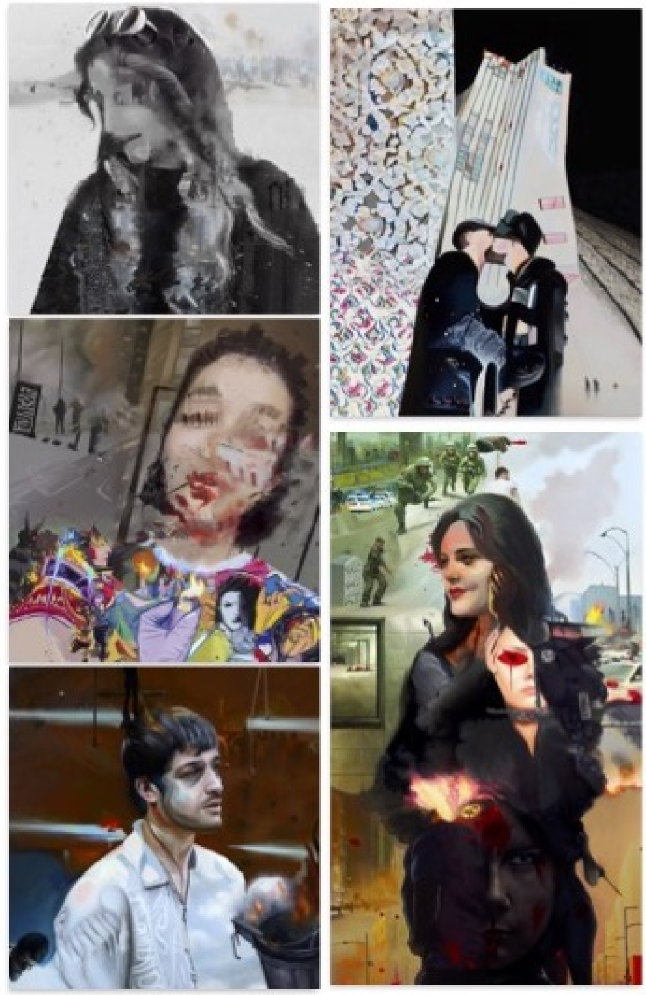


Figure 3: *Tulips of Freedom* collection

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. *CoRR* abs/2103.00020.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models.

Zylinska, J. 2020. *AI Art: Machine Visions and Warped Dreams*. Open Humanities Press.

Improving Efficiency and Coherence in Evolutionary Story Generation

Pablo Gervás
Facultad de Informática
Universidad Complutense de Madrid
Madrid, 28040, Spain

Abstract

A significant challenge for evolutionary approaches to story generation is to find a genetic representation for a story draft that allows mutation and crossover operations while also being able to capture the constraints of coherence between the different parts of the story. This may be achieved by defining a narrative draft in terms of combinations of knowledge structures that capture its structure. The present paper reviews a previously existing solution for the evolutionary generation of stories, both in terms of its representation, the evolutionary operators and the fitness function, and outlines an alternative solution that improves upon it. The two solutions are compared in terms of coverage of the search space, efficiency of the evolutionary search process, and quality of the resulting narratives.

Introduction

Evolutionary solutions have proven to be appropriate for implementing story generation systems based on models of desirable stories rather than on models of how humans build stories. This is because the essence of an evolutionary process lies in informed selection among a population of candidates, with the construction of the candidates being modelled on evolution via random mutation and crossover.

The challenge for applying this type of process to story generation arises from the choice of a genetic representation. A genetic representation for a story draft must allow mutation and crossover operations while also being able to capture the constraints of coherence between the different parts of the story. Solutions based on exclusively local representation of the different spans of the story will lead to outputs similar to those produced by the *exquisite corpse* technique of the Surrealists (Adamowicz 1998) or the *cut up* technique of the Dadaists (Cran 2013). Fragments cut out literally from different drafts will most often not make sense when put together in a new one. To avoid this problem, representations must be chosen that represent the structure of narrative in a way that captures its internal relations, but which, when some part of it is altered–as by evolutionary operators of mutation or cross over– it results in a different narrative that is also structurally coherent.

The present paper reviews a previously existing solution for the evolutionary generation of stories, both in terms of its representation, the evolutionary operators and the fitness function, and outlines an alternative solution that improves upon it. The two solutions are compared in terms of coverage of the search space, efficiency of the evolutionary search process, and quality of the resulting narratives.

Previous Work

The work presented in this paper requires understanding of three aspects of story generation: plot representation, prior evolutionary approaches and the existing approach used as starting point.

Basic Challenges of Plot Representation

Good stories have plot: the events in them are connected by a sense of causality (Forster 1927). Forster's famous argument states that "*The king died. The queen died.*" is a chronology of events, but "*The king died. The queen died of grief.*" is a plot. Knowledge-based procedures for story construction rely on capturing relations between events in some form in the representations they use for stories. Causal relations between events can be captured over complete *story schemas* (Booker 2004) or by defining smaller building blocks–such as *planning operators*–that define preconditions and postconditions with other elements in the story (Young et al. 2013). An intermediate approach relies on *axes of interest* or AoIs–small sequences of *plot atoms* representing events connected by plot-relevant causality and sharing characters in *roles* important to the plot (Gervás 2019). Table 1 shows an example of two AoIs combined into a simple plot.

Evolutionary Story Generation

Evolutionary algorithms have been applied to combine story fragments involving particular entities to the story, relying on a fitness function that combines coherence and interest of the story (McIntyre and Lapata 2010) or to generate small narrative fragments for text-based games using an evolutionary solution driven by novelty (Fredericks and DeVries 2021).

Other approaches have combined planning-based techniques to generate stories with evolutionary selection based on fitness functions. Aspects considered in the fitness functions are the believability of the story and the percentage of the user-defined goals the current story accomplishes (Kartal, Koenig, and Guy 2014) or degree of matching between

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

AoI	Plot Atom	Roles
ABDUCTION	Kidnapping	(abductor=x, abducted=y)
	Rescue	(abducted= , rescuer=z)
CALLTOACTION	Call	(called=hero,caller=sender)
	Reward	(rewarded=x)

(a) two axes of Interest (AoIs)

AB	Kidnapping(abductor=*villain*,abducted=*victim*)
CA	Call(called=***hero***,caller=*sender*)
AB	Rescue(abducted=*victim*,rescuer=***hero***)
CA	Reward(rewarded=***hero***)

(b) a combination of them into a simple plot (protagonist in Bold, rest of the characters in Italic).

Table 1: Plot representation in terms of AoIs.

AoIs	Abduction	(relation)	CallToAction
Shared roles	hero	=	hero
Sequencing	Abduction	<	CallToAction
	Rescue	>	CallToAction
	Rescue	<	Reward

Table 2: Example of constraint: the hero of both AoIs must be the same (line 2), the abduction must take place before the call to action (line 3), the rescue must take place after the call to action (line 4) and before the reward (line 5).

the tensions in the story and a target curve of evolving tensions provided as input (de Lima, Feijó, and Furtado 2019).

Our Starting Point

The evolutionary solution in (Gervás 2022) combines AoIs (see Table 1 above) using as fitness function the correct sequencing of events and acceptable occurrence of characters sharing roles across AoIs. The genetic representation employed for evolutionary construction of stories represents a narrative in terms of how the plot atoms in the AoIs are presented in the ordered sequence that constitutes the discourse of the narrative, and how the various roles for characters in the plot atoms are instantiated with identifiers for the characters in the narrative.

The fitness function that drives the evolutionary process relies on metrics for sequencing of events, and occurrence of characters sharing roles across AoIs proposed in (Gervás 2022). For each pairwise combination of AoIs the constraints on character occurrence and event sequencing are expressed in the form of constraints as shown in Table 2. The metrics assign a partial score over 100 to each sequencing constraint over events, corresponding to the number of positions that one of the elements would need to shift for the constraint to hold (normalised over the length of the sequence). Each role-sharing constraint present is scored 100 if met and 0 otherwise. The final score for a draft is computed as the weighted sum of the average value of the role-sharing constraints and the average value of the sequencing constraints. The relative weights for sequencing and role sharing constraints have been empirically set to 20 and 80.

This metric provides a progressive scoring, so that drafts where the sequencing constraints are not met are scored relative to how far they need to be modified for the constraints to be met. This allows mutations that modify the sequence in the right direction to be scored progressively higher, allowing evolution to converge towards optimal solutions.

Optimising the Evolutionary Process

A detailed study of the performance of the original algorithm lead to the identification of some shortcomings, which, when solved, lead to significant improvements in performance.

Issues with the Original Genetic Representation

The original evolutionary solution relied on a genetic representation that presented three important shortcomings. First, it represented the order in which the plot atoms appeared in the story in terms of the set of jumps to be made over the the constituent AoIs. Small changes in the set of jumps lead to very different final stories. This lead to poor exploration of the search space, because it made it difficult to explore alternatives close in the neighbourhood of given individual. Second, the genetic representation allowed jumps to be postulated even when there were no further AoIs available to jump to, having all been exhausted in prior jumps. This created situations in which different genetic representations–one that indicated a shift to another AoI at that point and one that did not–resulting in the same actual narrative. This had a negative side effect in that populations might have individuals with different genotype but equivalent phenotype. Third, the representation for the instantiations of roles from different AoIs with shared characters lead to assymmetries between different parts of the encoding vector: positions at the start of the vector had a wealth of potential candidates to be instantiated, whereas positions later in the vector could only be instantiated with characters already assigned to incompatible positions earlier. This also lead to underperformance of the evolutionary algorithm when exploring the search space.

These shortcomings went unnoticed in the early tests because it was assumed that the observed low scores were the result of incompatible restrictions for a given set of AoIs. However, more detailed consideration lead to the discovery of the negative impact of these problems in the genetic representation, which were stopping the evolutionary algorithm from reaching more desirable areas of the search space.

An Improved Genetic Representation

The original genetic representation has been replaced with a new version that resolves the observed shortcomings. It still encodes separately the order in which the plot atoms from the various AoIs appeared in the discourse and the instantiations of roles from different AoIs with shared characters of the story.

The order of appearance is now encoded as sequence of indices of the plot atoms to be included in the discourse. Each index simply indicates which plot atom from which AoI should feature next in the discourse. Mutation is now encoded as a shift of a particular index either forwards or backwards in the sequence for a number of positions chosen at random. Shifts involve skipping over plot atoms from other AoIs but they must respect the relative order of plot atoms within the same AoI. This encoding does not allow crossover operations, as cutting different representations at the same point is likely to result in drafts with either missing or redundant instances of plot atoms in some AoIs.

The instantiations of roles from different AoIs with shared characters of the story is now encoded as a set of specific data structures for encoding any variables that have a shared instantiation across pairs of AoIs. Mutation is now encoded as either adding or eliminating a connection to the data structure for a particular pair of AoIs. The choice of which pair of AoIs to consider and whether to add or eliminate are chosen at random within the bounds of available possibilities. Pairs with no connections only allow addition, pairs with all available variables already connected only allow elimination.

Although the new representation no longer allows crossover operations on the subsets of the genetic representation that encode the different aspects, a certain crossover is possible by swapping the representations of the relative order in the sequence between two different individuals to give rise to a new pair.

An example of system output encoded with the new genetic representation is shown in Table 4. This example shows together the improved genetic representation–the genotype– and the instantiation of it as story–the phenotype. Additionally it shows the intermediate data structures that translate the genetic encoding into the features that are used to construct the final draft for the narrative.

The textual rendering presented for the narrative is not intended to be the final medium for presenting it to a potential audience. Since the generation procedure described here is only concerned with the narrative structure of the plot, it is beyond the scope of the paper to evaluate or even consider aspects specific to the linguistic rendering of this content. Nevertheless a template-based transcription of the content is included to facilitate the appreciation of the narrative structure. Alternative solutions based on neural technologies, such as generative pre-trained transformers (Dale 2021), may be considered in future work.

Metric for Romantic Coherence

An undesirable feature of the early results was the fact that the resulting narratives exhibited in some cases incoherent behaviour of the characters in terms of their romantic inclinations. As many of the AoIs involve romantic relations between the characters, this often resulted in stories where characters exhibited surprisingly promiscuous behaviour, such as marrying several different characters in succession with on intervening explanation of their change of heart. These situations came about when two AoIs were combined that both included romantic relations between the characters, for instance SHIFTINGLOVE–which involves a character oscilating between two different love interests as the story evolves and deciding on one towards the end–and RELENTINGGUARDIAN–which involves a couple who wants to marry overcoming the obstacle of a guardian opposed to the match. When these two AoIs are combined, there were no safeguards in the original solution to avoid that a single character end up being matched with two different partners, one under each AoI.

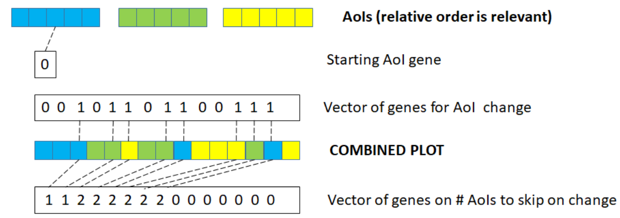


Figure 1: Original genetic representation.

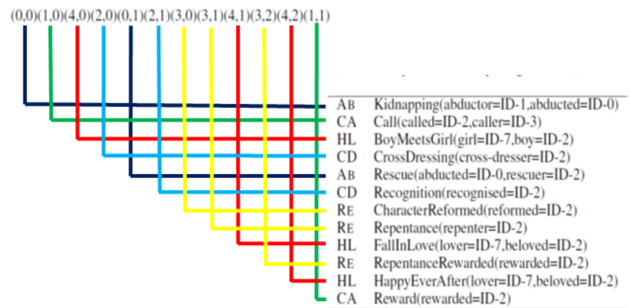


Figure 2: Improved genetic representation.

To filter out these cases, an additional component was added to the metrics that scored each character in a draft in terms of the their romantic consistency. Each character in a draft is now assigned a romantic consistency score of 100 if has at most one single romantic match, and each draft is assigned the average of the scores on romantic consistency of the characters in it. This additional metric is added to the existing fitness function, which is already computed as an average of a number of metrics on consistency over different pairs of AoIs.

Discussion

The two genetic representations considered for the discourse sequence of the story are shown in Figures 1 and 2. The original representation encoded the operations required to combine the AoIs, including a gene to indicate which AoI to start the story on, binary genes for each position in the draft to indicate whether a transition to a different AoI followed, and numerical genes to indicate how many AoIs to skip in each transition. In contrast, the new representation encodes simply the final order of the discourse.

The introduction of the new representation has lead to a significant increase in the average scores of the populations for runs on equivalent sets of inputs. The results may be compared in terms of the relative scores on quality, because both the knowledge resources and the evaluation metrics used as fitness functions remain the same. A quantitative comparison of the score for the two versions is shown in Table 3.

Whereas the results of the earlier version converged to the scores considerably below the maximum threshold, the current version reaches achieves consistently higher scores under similar circumstances. It also does frequently reach the

Version	Pop.size	Average score	Highest score
Original	100	59.2	80.1
Improved	100	67.0	94.6
Original	200	59.1	77.9
Improved	200	73.3	98.0

Table 3: Comparative scores for the evolutionary solution based on the original genetic representation and the improved version. Scores shown are averages over 10 runs of each system with the same setting: population size of 100 and 200, 20 generations, seed RAGS2RICHES, expanding with 3 additional AoIs.

maximum score. This is due to a significantly better exploration of the search space. The earlier version of the procedure must have been stuck in local optima held back by redundant encoding. This hypothesis is supported by the fact that an increase in the size of the population does not yield any significant changes in the scores. It is important to note that, while the increase in the size of the population does lead to a slight increase in the scores for the version using the improved representation, the scores for the version using the original representation are even lower than with a smaller population.

The shortcomings observed in the original genetic representation correspond to the problems of non-synonymous redundancy and low locality (Rothlauf 2006). The improved performance of the proposed solution highlights the importance for evolutionary approaches to story generation to satisfy such general requirements on genetic representations.

This improvement in the overall scores that arises from the modified genetic representation is compensated by the introduction of the additional metric on romantic consistency. With this addition, the final scores of the population recover their discriminating capability, and the results now include narratives that are coherent with respect to the romantic lives of the characters. The metric for romantic coherence is consistent with prior approaches to evaluating semantic coherence of significant events over story drafts (Gervás, Concepción, and Méndez 2022).

In general terms, the set of metrics integrated into the fitness function are defined over characteristics that are specific to the phenotype rather than the genotype of each draft. For this reason, they are applicable to any stories regardless of whether they have been produced by an evolutionary procedure or any other construction method. To apply these metrics to stories beyond the output of this system the only requirement is to provide means for the specific features being considered–narrative roles, plot-relevant events, story milestones that imply romantic commitments. . . –to be extracted from the stories to be considered. As the significance of such features for the evaluation of stories is difficult to question, the set of metrics in themselves can be considered a valuable contribution to the field.

This is especially useful in a context where the application of neural technologies has lead to a proliferation of solutions for story generation based exclusively on probabilities of word co-occurrence. Such solutions are known to be susceptible of significant improvement by means of fine tuning procedures driven by reinforcement learning based on computational reward models (Ziegler et al. 2019). If the problem of automatically extracting semantic information from text can be solved successfully, metrics such as these can prove to be valuable contributions for solutions based on large language models as support either for reward models during fine tuning or for filtering and refining outputs–in the processes known as prompt engineering.

Conclusions

The improved genetic representation proposed for evolutionary combination of plot-relevant spans of discourse solves the shortcomings observed in prior versions. The stories obtained with the enhanced version achieve significantly higher scores under the existing metrics for story quality, leading to a point where system outcomes consistently reach top scores.

As this endangers the discriminating power of the metrics for identifying higher quality stories, an extension of the metric has been proposed. The extension measures the coherence of the romantic behaviour of the characters. Under the extended set of metrics the system generates stories that have recognisable narrative structure and in which the characters are consistent in terms of their romantic relations.

The metrics on story quality proposed to inform the evolutionary fitness functions are designed to captures features relevant to the evaluation of narrative and they are independent of the genetic representation and the overall evolutionary procedure. They are therefore valuable contributions to the field of story generation in general on their own right.

As future work we intend to address issues at two different levels. In terms of richer representations of narrative structure, we will explore extensions to the construction procedure to make it capable of generating narratives as a series of connected episodes. In terms of improvements on the rendering of the narratives as text we will consider solutions for rendering the resulting plots as text that rely on generative pretrained transformers.

Acknowledgments

This paper has been partially funded by the projects CANTOR: Automated Composition of Personal Narratives as an aid for Occupational Therapy based on Reminescence, Grant. No. PID2019-108927RB-I00 (Spanish Ministry of Science and Innovation) and the ADARVE (Análisis de Datos de Realidad Virtual para Emergencias Radiológicas) Project funded by the Spanish Consejo de Seguridad Nuclear (CSN).

References

Adamowicz, E. 1998. *Surrealist collage in text and image: Dissecting the exquisite corpse*, volume 56. Cambridge University Press.

Booker, C. 2004. *The Seven Basic Plots: Why We Tell Stories*. The Seven Basic Plots: Why We Tell Stories. Continuum.

(a) genetic representation for discourse order: each pair shows (<AoI index>,<plot atom index>).

(0,0)(1,0)(4,0)(2,0)(0,1)(2,1)(3,0)(3,1)(4,1)(3,2)(4,2)(1,1)

(b) genetic representation for character instantiations: each column represents a pair of AoIs, row 1 holds indices of the AoIs, row 2 shows how many variables appear in each AoI, row 3 shows which variable from the first is connected to which of the second.

0-1	0-2	0-3	0-4	1-2	1-3	1-4	2-3	2-4	3-4
3-2	3-1	3-1	3-2	2-1	2-1	2-2	1-1	1-2	1-2
2-1			0-1	1-0	0-0		0-0	0-1	

(c) discourse plan encoded by the genes in (a).

AB	Kidnapping(abductor=ID-1,abducted=ID-0)
CA	Call(called=ID-2,caller=ID-3)
HL	BoyMeetsGirl(girl=ID-7,boy=ID-2)
CD	CrossDressing(cross-dresser=ID-2)
AB	Rescue(abducted=ID-0,rescuer=ID-2)
CD	Recognition(recognised=ID-2)
RE	CharacterReformed(reformed=ID-2)
RE	Repentance(repenter=ID-2)
HL	FallInLove(lover=ID-7,beloved=ID-2)
RE	RepentanceRewarded(rewarded=ID-2)
HL	HappyEverAfter(lover=ID-7,beloved=ID-2)
CA	Reward(rewarded=ID-2)

(d) character assignment encoded by the genes in (b).

Abduction-hero	ID-2
Abduction-victim	ID-0
Abduction-villain	ID-1
CallToAction-sender	ID-3
HappyLove-girl	ID-7
HappyLove-hero	ID-2
CrossDressing-someone	ID-2
CallToAction-hero	ID-2
Repentance-villain	ID-2

(e) template-based text rendering of the plot of the draft.

Scott kidnaps Hawa. Korr calls to action West. West meets Lilly. West dresses up as a member of the opposite sex. West rescues Hawa from Scott. West is recognised. West reforms their character. West repents. West falls in love with Lilly. West sees repentance rewarded. West lives happily ever after with Lilly. West is rewarded.

Table 4: An example of plot generated by the system by combining the following AoIs: ABDUCTION (AB), CALLTOACTION (CA), CROSSDRESSING (CD), REPENTANCE (RE), and HAPPYLOVE (HL). The procedure was initiated using ABDUCTION as a seed, to be expanded with 4 additional AoIs. The evolutionary process was run for 30 generations with a population of 500 individuals.

Cran, R. 2013. 'Everything is permitted': William Burroughs' Cut-up Novels and European Art. *Comparative American Studies An International Journal* 11(3):300–313.

Dale, R. 2021. Gpt-3: What's it good for? *Natural Language Engineering* 27(1):113–118.

de Lima, E. S.; Feijó, B.; and Furtado, A. L. 2019. Procedural generation of quests for games using genetic algorithms and automated planning. In *18th Brazilian Symposium on Computer Games and Digital Entertainment, SBGames 2019, Rio de Janeiro, Brazil, October 28-31, 2019*, 144–153. IEEE.

Forster, E. M. 1927. *Aspects of the novel*. New York: Harcourt.

Fredericks, E. M., and DeVries, B. 2021. (Genetically) Improving Novelty in Procedural Story Generation. In *2021 IEEE/ACM International Workshop on Genetic Improvement (GI)*, 39–40. IEEE.

Gervás, P.; Concepción, E.; and Méndez, G. 2022. Evolutionary construction of stories that combine several plot lines. In *Computational Intelligence in Music, Sound, Art and Design – 11th International Conference, EvoMUSART 2022*. Madrid, Spain: Springer.

Gervás, P. 2019. Generating a search space of acceptable narrative plots. In *10th International Conference on Computational Creativity (ICCC 2019)*.

Gervás, P. 2022. Evolutionary stitching of plot units with character threads. In *WIVACE 2022 XVI International Workshop on Artificial Life and Evolutionary Computation*.

Kartal, B.; Koenig, J.; and Guy, S. J. 2014. User-driven narrative variation in large story domains using monte carlo tree search. In *Procs. of AAMAS '14*, 69–76.

McIntyre, N., and Lapata, M. 2010. Plot induction and evolutionary search for story generation. In *Procs. of ACL 2010*, 1562–1572. Uppsala, Sweden: Association for Computational Linguistics.

Rothlauf, F. 2006. *Representations for genetic and evolutionary algorithms*. Springer.

Young, R. M.; Ware, S. G.; Cassell, B. A.; and Robertson, J. 2013. Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung* 37(1-2):41–64.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P. F.; and Irving, G. 2019. Finetuning language models from human preferences. *CoRR* abs/1909.08593.

Using ChatGPT for Story Sifting in Narrative Generation

Gonzalo Méndez and **Pablo Gervás**
Facultad de Informática
Universidad Complutense de Madrid
Madrid, Spain
{gmendez,pgervas}@ucm.es

Abstract

The task of selecting a subset of story-worthy events from out of an observed collection of facts–known as *story sifting*–is a useful human ability that has yet to be emulated successfully by computational processes. The emergence of Large Language Models (LLMs) has made it necessary to rethink the way of carrying out many tasks that were previously performed using other tools. This short paper explores how the infamous ChatGPT fares when asked to sift stories from the log of an agent-based simulation featuring romantic relations between characters.

Introduction

In an average day we experience or observe a multitude of events that register in our consciousness, yet at any given point any one of us is capable of isolating a small subset of those events as being appropriate for piecing together into an interesting story to tell about our day. The successful computational modeling of such processes is at this point an open question. This task, known as *story sifting*, is being successfully addressed by other less glamorous AI techniques.

The coverage in the media of the recent triumphs of generative AI based on large language models has created the impression that they can be successful at any task. Therefore it is particularly important that they be evaluated scientifically to establish the extent of their capabilities and, if any, of their limitations. The present paper explores the applicability of one of these models to the task of extracting from a set of events a subset that, when presented in narrative form, results in a story of higher quality than the set of events presented wholesale.

Previous Work

We outline some background on the story sifting task and some basic characteristics of ChatGPT.

Story Sifting

Early work on narrative generation produced literary texts by selecting a subset of lines from an extensive source file (Montfort and Fedorova 2012). A subsequent refinement on this technique mines sequences of events corresponding to interesting stories from the logs of agent-based simulations. James Ryan's PhD thesis (Ryan 2018) outlines how, rather than automatically inventing stories, narrative may emerge from the activity of characters set in motion in a simulated story world, and defines the task of curating such narratives out of simulation logs as story sifting. The Felt story sifting and simulation engine (Kreminski, Dickinson, and Wardrip-Fruin 2019) introduced the concept of *story sifting patterns*, which are descriptions of sequences of events that exhibit high potential to be part of interesting narratives. He develops tools for authoring such patterns and applying them to sets of events to implement automated story sifting. This line of research lead to the development of Winnow (Kreminski, Dickinson, and Mateas 2021), a domain-specific language for specifying story sifting patterns that can be run on ongoing simulations to identify event sequences with narrative potential.

As it can be seen, it is necessary to consider how well this task can be performed using LLMs, what are their possible limitations and in which cases it is worthwhile to continue using other types of techniques. Although in this contribution we do not intend to answer these questions extensively, we provide our point of view, based on our experience, on how well one of these language models, ChatGPT[1], performs this task.

ChatGPT

ChatGPT is an interactive online system that responds to textual prompts presented by the user with fluent prose that always appears to be an acceptable response to the given prompt. Although there is no recognised scientific publication that describes how ChatGPT operates, it is public knowledge that it is a member of the generative pre-trained transformer (GPT) family of language models (Radford et al. 2018), fine-tuned using reinforcement learning (MacGlashan et al. 2017; Ziegler et al. 2019). By virtue of this, it combines the advantage of a neural representation as semantics (Levy and Goldberg 2014), the linguistic fluency of transformers (Khan et al. 2022) and the knack of finding appropriate responses associated with reinforcement learn-

[1](`https:\\chat.openai.com`) ChatGPT Free Research Preview, Mar 23 Version (GPT 3.5)

ing. This allows it not only to respond fluently to most user requests, but also to understand clarifications or corrections and respond by adapting its prior outputs as requested by the user. It is this particular functionality that we hoped to explore in this paper for the story sifting task.

Story Sifting with ChatGPT

Story sifting involves processing an input that includes a large set of events pertaining to a specific story world to select a particular subset of those events such that, presented in a narrative fashion, they exhibit narrative qualities beyond those of the whole or any summary of it.

Input Simulation Log

We have tested the performance of ChatGPT in the story sifting task using a simulation log created by an agent-based simulation of characters that interact with each other in a romantic setting. In this simulation, characters can invite others to participate in different events together in order to increase their mutual affinity levels. The aspect of such interactions in the log is the following:

```
Suzette PROPOSE friend_chat Silvy
Silvy ACCEPT-PROPOSAL friend_chat Suzette
Silvy affinity with Suzette 57
Suzette affinity with Silvy 74
[...]
Betty PROPOSE mate_invite_dinner Robert
Robert REJECT-PROPOSAL mate_invite_dinner Betty
Robert affinity with Betty 93
Betty affinity with Robert 79
```

The log we have worked with was 2500 lines long and involved mixed interactions between 15 characters. The following types of events are considered: `PROPOSE`, `ACCEPT-PROPOSAL` and `REJECT-PROPOSAL`. In each case the entry starts with the active agent, and the activity in question and the potential participant follow. The affinity value between characters is not symmetrical and ranges from 0 to 100, which makes the affinity levels take the values of mate, friend, indifferent or foe.

Getting ChatGPT to Accept the Log

The first goal was to let ChatGPT read the log file and interpret it in order to start carrying out the story sifting task, and this is where we encountered the first obstacle, as the version we used only allows interaction by means of the chat, which involved pasting the content of the log file directly in the chat. Following ChatGPTs instructions, we started by pasting the full log in the chat, which resulted in an error. After several attempts reducing the size of the pasted text we ended up working with batches of 100 lines, as bigger sizes produced errors or involved a very poor performance.

The first batch with this size was correctly read and interpreted, and ChatGPT provided the following output

> Based on the log you provided, it seems to be a record of social interactions between different individuals. The log includes information about proposals, acceptances, and rejections of social activities such as going out, having dinner, and going to the cinema. It also includes information about the affinity level between the individuals involved in each activity. Additionally, it includes some information about negative interactions between some individuals, such as insults and ignoring. It is possible that this log is part of a simulation or game that simulates social interactions between individuals.

However, we encountered a new obstacle after we entered the next few batches, as ChatGPT seems to forget its own instructions and provides the following output:

> It seems like a log of interactions between several people. Do you have any specific question or request related to this log?

A couple of inputs later, it decided we wanted a summary of the events in the log. We tried to circumvent these limitations in different ways (e.g trying to tell ChatGPT the number of batches we were going to paste and numbering all of them) but ChatGPT seems to forget the instructions after very few interactions (usually three). Eventually, we were forced to work with only one set of 100 lines, as ChatGPT not only forgot the instructions, but also ignored sequential inputs and tended to work with the last batch we pasted. At this point, we identified this issue as a limitation in the extent of our experiment and selected the first 100 lines of the log in order to start with the story sifting process.

Story Sifting Tests

Having reached the point where the ChatGPT system might be considered aware of the content of the log to be processed, we presented the following prompt to request that it attempt to generate a story from the log, so we could have a baseline of what we could expect next:

> Can you write a story using this log information?

ChatGPT's response is reproduced below:

> Once upon a time, in the year 2022, a group of friends had a complex social dynamic that was closely monitored by an AI system that tracked their proposals and affinities. Megan proposed to have lunch with Meredith, and the latter accepted. Meanwhile, Lester proposed to chat with Robert, but Robert rejected the proposal. Suzette proposed to chat with Silvy, and the latter accepted. Betty proposed to go out on a weekend with Clark, but Clark rejected the proposal. Meredith proposed to watch TV with Lester, but he rejected the proposal.
>
> (The story expands for 5 similar paragraphs)

The story, as it can be observed, is only a little more than a direct translation from the log format into text, almost lacking any narrative quality.

Next, we asked ChatGPT to write a story using only the most significant events (which would be a general description of what story sifting consists in). The result in this case was somehow similar to the previous one, a story that omitted some potentially relevant events, specifically all the breakups that appear in the log, as they result in drastic changes in the affinities between characters. It first claimed that there were no breakups in the log, but later acknowledged that it had "missed" the events.

When asked why and how it had decided which were the most relevant events, the answer was:

> I determined that these events were the most relevant based on the frequency and variety of the actions that took place, as well as the different relationships between the characters that they reveal. [...] The events I mentioned are some of the most frequent and varied actions that occur in the world and are therefore important to be aware of.

Further attempts to get more elaborated and specific responses did not shed any light to the previous explanation. Notably, the changes in the affinities between characters were deliberately ignored.

After this, we tried to highlight the importance of the affinity values by asking ChatGPT to write a story using the pairs of characters that experienced larger variations in their mutual affinity values, but the result was similar to the previous case, obtaining a story that omitted all breakups between couples.

Watching these results, we decided to be more specific about the intended output, so the following input directed the generated result towards one of the couples who brake up:

> Can you write the story of a couple that breaks up?

ChatGPT's response is partially reproduced below:

> Once upon a time, Megan and Lester were a happy couple. They had been together for several years and had built a life together. They enjoyed spending time together, going on adventures, and making memories. However, as time went on, they began to grow apart. They stopped communicating effectively and stopped making an effort in the relationship. Lester began to withdraw and stopped showing Megan affection. He would spend more time at work or with his friends, leaving Megan feeling alone and neglected.
>
> One day, Megan met Meredith and they hit it off right away. They began spending more time together, going on lunch dates and watching TV together. Lester became jealous of the time Megan was spending with Meredith and rejected her proposal to watch TV together.
>
> Meanwhile, Clark rejected Betty's proposal to go out for the weekend, leaving Betty feeling rejected and unwanted. Suzette and Silvy hit it off and began chatting regularly. Robert rejected Lester's proposal to chat, causing Lester to feel rejected and unwanted.
>
> [...]

As it can be seen, ChatGPT makes a big effort to provide a background to the story, but it takes too short to start rambling and including events about characters that do not take part in the requested narrative thread, therefore producing a garbled result.

An even more focused attempt to obtain a specific output for the story sifting process consisted in specifying the characters that were to be considered in the narrative, with the following result:

> Can you tell a story of Betty and Clark using the information in this log?

ChatGPT's response is partially reproduced below:

> Betty and Clark had been friends for a while, and Betty had been hoping to spend some more time with him. One day, she proposed that they go out together for the weekend. Unfortunately, Clark rejected her proposal. Although she was disappointed, Betty tried to move on.
>
> Meanwhile, Meredith proposed to Lester that they watch TV together as mates, but he rejected her proposal. Violet also proposed to Megan that they chat, but Megan rejected her proposal.
>
> Megan proposed to Tony that she give him a massage, but he also rejected her proposal. However, they still had a high affinity score, indicating that they still liked each other as friends.
>
> [...]

Again, ChatGPT decided to include other events that are present in the log instead of complying with our request. At this point, instead of trying to force a specific output, we tried to give ChatGPT a little more freedom by only restricting the number of characters that should be considered:

> Can you select the two most interesting characters from this log and tell their story?

ChatGPT's response is partially reproduced below:

> Meredith and Violet were once best friends. They did everything together, and everyone knew that they were inseparable. However, things took a turn when Meredith proposed to have lunch with Megan, and Violet proposed to chat with her. Meredith accepted Megan's proposal, but rejected Violet's proposal. This made Violet upset, and she proposed to hug Megan instead. However, Megan rejected Violet's proposal.
>
> As time passed, Meredith and Violet's relationship deteriorated. They had a falling out, and Meredith even informed others that she despised Violet. She also accused Violet of cheating on her, which further strained their relationship. Meanwhile, Meredith developed an affinity for Lester, while Violet's affinity with Meredith dwindled.
>
> [...]

Apart from a small inconsistency in the fact that Meredith proposes Megan to have lunch but then it is Meredith, and not Megan, who accepts the proposal, reading the log it is remarkable to see that the facts in the first paragraph do not appear in it or appear in a different way. Additionally, when asked why Meredith and Violet are the most interesting characters in the log the answer is:

> Overall, the interactions and dynamics between these characters have the potential to create a compelling story.

This means that it has selected the characters not only for what they have done, but for what they might do in the rest of the log according to what ChatGPT expects from what it has seen. Another plausible interpretation is that, provided that ChatGPT basically relies on a LLM, the explanation provided by ChatGPT is a carefully crafted text that does not really reflect what is happening underneath.

At this point, we decided to stop out experiments using ChatGPT for story sifting, as it was already patent that ChatGPT has a high potential to create unrestricted narratives, but it lacks ability to perform tasks such as story sifting, which require sticking to a plan in a more consistent way.

Observed Limitations

The following limitations have been observed during the experiments carried out:

- ChatGPT seems to forget the instructions after very few interactions, which limits the possibility of working with a big amount of data
- ChatGPT ignores sequential inputs and tends to work with the last batch received
- ChatGPT 's response to a set of events appears to be restricted to summarising it, very neatly but with no option of selecting a particular subset based on its potential narrative qualities
- ChatGPT's output is quite biassed towards adding its own content, thus making it difficult to restrict the result to the data provided as input
- ChatGPT's criteria to perform story sifting operations are vaguely explainable, generic and difficult to influence in order to obtain results that follow different criteria or to adjust the to a specific domain

As a result, our impression is that there is still room for improvement in relation to LLM-based story sifting, and consequently there is still need to keep on using and researching on other techniques for this task.

Comparison with an Evolutionary Sifting Method

In order to make our position stronger and show that it is possible to outperform ChatGPT in the story sifting task, we have used the original log file containing 2500 lines as an input for a story sifting algorithm specifically crafted for this task. It has been developed using a genetic algorithm that makes use of heuristics based on the affinity values in order to optimize the story sifting process. The output of this algorithm is partially reproduced bellow:

ProposeActivity-544 mate_go_to_cinema Lester Meredith
ActivityRejected-545 mate_go_to_cinema Lester Meredith
ProposeActivity-629 mate_dinner_with_candles Lester Meredith
ActivityRejected-631 mate_dinner_with_candles Lester Meredith
ProposeActivity-747 mate_hug Lester Meredith
ActivityRejected-748 mate_hug Lester Meredith
ProposeActivity-975 mate_watch_tv Lester Meredith
ActivityAccepted-976 mate_watch_tv Lester Meredith

In this case, we acknowledge the superior quality of the texts generated by Large Language Models, and more specifically ChatGPT, over other techniques, so we have used it in order to generate the final (partial) story, as shown bellow:

> Lester had been trying to plan the perfect date for weeks. He wanted something special, something that would show Meredith just how much he cared about her. But every idea he proposed seemed to fall flat.
>
> First, he suggested they go to the cinema to see a romantic movie. Meredith wasn't interested and rejected the idea. Then, he proposed a dinner with candles, hoping to create a romantic atmosphere. But again, Meredith rejected the idea.
>
> Feeling dejected, Lester decided to try something simpler. He proposed they just spend some quality time together watching TV. Meredith accepted, and they settled in on the couch with some snacks and drinks.
>
> As they watched their favorite show, Lester put his arm around Meredith and gave her a hug. She leaned into him, and they both felt the warmth and comfort of each other's embrace.
>
> In that moment, Lester realized that it wasn't about the grand gestures or elaborate plans. It was about the simple moments of connection that made all the difference. And as they sat there together, watching TV and cuddling, he knew that he had found something special with Meredith.

It is impressive to see how ChatGPT has managed to expand the bare sketch of events provided in the log into a moving story. The elements sifted from the log as pertaining to a potentially interesting story have been expanded into paragraphs that enrich the story with detail and nuance. Even so, the example provided shows an ideal output generated by ChatGPT, as previous attempts also included events that were kindly provided by it, but which were not present in the output generated by the genetic algorithm. It is clear that solutions like ChatGPT have a significant potential in terms of what they contribute to story generation. However, they still have room for improvement in terms of specific tasks such as story sifting.

In order to compare both models in similar conditions, in Table 1 we show the selection made by each of them over a log of 100 lines.

Potential Applications

It is clear from the reported results that the evolutionary story sifting process achieves better results in terms of identifying a relevant subset of events from a given log but presents a poor textual rendition. At the same time, it is clear that ChatGPT has difficulty in grasping the concept of "selecting a relevant subset of events" and inserts events that were not present in the log, but generates a text that is impressive by its fluency. These two observations suggest that the optimal solution might be found in an informed combination of the two approaches. In such a combination the evolutionary sifting algorithm might be relied upon to carry out the selection of the relevant events and ChatGPT to render the resulting selection as text. Any such combination may need to be refined to ensure that any events hallucinated by the neural solutions are identified and filtered out.

Potential applications of such models would include automated pre-processing of system logs or surveillance records to highlight relevant sequences of events, or automated generation of relevant narrative threads in video games.

Conclusions

The experiments carried out show that it is at present difficult to get ChatGPT to carry out processes of story sifting from simulation logs, as this is a knowledge related task more than a simply language related one. Although the prose of the responses is fluent and sounds natural, the concept of focusing on narrative threads that restrict the narrative to particular subsets of the characters appears to be beyond the current capabilities of the system. It is possible that fine-tuning the underlying language model with a story-sifting specific dataset may improve the results produced by these models, but this is something that still needs to be researched.

Successive efforts of prompt engineering angling for the appropriate responses have not found the expected result. This is in spite of the impressive ability of the system to come up with appropriate responses to most requests.

The results obtained generally constitute valid summaries of the material provided, often presented in reasonable narrative form and in fluent prose. However, they tend not to satisfy the requirements for a valid process of story sifting, which are different from those of simple summarisation. Whereas summarisation involves finding a significantly shorter rendering of pretty much the same material in the input, story sifting should involve a process of deciding to focus on a subset of the material such that the narrative quality of the result is significantly higher than the original input or any summary of it. This concept appears to present difficulties to ChatGPT.

We nevertheless believe that there is significant potential in the idea of applying neural solutions to the story sifting task.

As further work, we intend to explore the use of alternative neural language models to check if they present similar limitations and propose approaches to overcome them.

Action	Evo	Chat
Megan PROPOSE have lunch Meredith		X
Lester PROPOSE chat Robert		X
Suzette PROPOSE chat Silvy		X
Betty PROPOSE weekend out Clark	X	X
Meredith PROPOSE watch tv Lester		X
Clark REJECT weekend out Betty	X	X
Lester REJECT watch tv Meredith		X
Meredith ACCEPT have lunch Megan		X
Violet PROPOSE chat Megan	X	X
Robert REJECT chat Lester		X
Silvy ACCEPT chat Suzette		X
Clark PROPOSE hug Betty		X
Betty PROPOSE invite dinner Robert		
Betty REJECT hug Clark		X
Robert REJECT invite dinner Betty		
Megan PROPOSE give massage Tony		X
Lester PROPOSE sleep together Meredith		
Mary PROPOSE play tennis Megan		X
Silvy ACCEPT chat Suzette		X
Meredith INFORM despise Violet		X
Tony REJECT give massage Megan		X
Megan REJECT chat Violet	X	X
Violet PROPOSE hug Megan		
Meredith INFORM cheat Violet		X
Megan PROPOSE go out John		
Betty INFORM break up Robert		
Mary PROPOSE help Silvy		
Suzette INFORM slander Mary		
Clark PROPOSE hug Ray		
Ray ACCEPT hug Clark		
Meredith REJECT sleep together Lester		
Betty PROPOSE have lunch Clark	X	
Clark ACCEPT have lunch Betty	X	
Suzette PROPOSE hug Silvy		
Silvy PROPOSE weekend together Drew		
Silvy ACCEPT hug Suzette		
Drew PROPOSE help Simon		
Silvy INFORM cheat Suzette		
Mary PROPOSE help Silvy		
Silvy REJECT help Mary		
Suzette PROPOSE have coffe Silvy		
Silvy PROPOSE talk Drew		
Silvy ACCEPT have coffe Suzette		
Drew ACCEPT talk Silvy		
Mary PROPOSE chat Megan		

Table 1: Comparison between a story sifting operation performed by the evolutionary algorithm and ChatGPT. The columns marked with X show the actions selected by each of them over a log of 100 lines (lines with changes in affinity levels have been removed for clarity and space reasons). The evolutionary algorithm tends to take into account all the events, while ChatGPT tends to focus more on the initial events

Acknowledgments

This paper has been partially funded by the projects CANTOR: Automated Composition of Personal Narratives as an aid for Occupational Therapy based on Reminescence, Grant No. PID2019-108927RB-I00 (Spanish Ministry of Science and Innovation); and ADARVE: Análisis de Datos de Realidad Virtual para Emergencias Radiológicas, Grant No. SUBV-20/2021, funded by the Spanish Consejo de Seguridad Nuclear (CSN).

References

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54(10s):1–41.

Kreminski, M.; Dickinson, M.; and Mateas, M. 2021. Winnow: A domain-specific language for incremental story sifting. *Proc. of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 17(1):156–163.

Kreminski, M.; Dickinson, M.; and Wardrip-Fruin, N. 2019. Felt: A simple story sifter. In *12th International Conference on Interactive Digital Storytelling, ICIDS 2019*, 267–281. Berlin, Heidelberg: Springer-Verlag.

Levy, O., and Goldberg, Y. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308.

MacGlashan, J.; Ho, M. K.; Loftin, R.; Peng, B.; Wang, G.; Roberts, D. L.; Taylor, M. E.; and Littman, M. L. 2017. Interactive learning from policy-dependent human feedback. In *International Conference on Machine Learning*, 2285–2294. PMLR.

Montfort, N., and Fedorova, N. 2012. Small-scale systems and computational creativity. In *International conference on computational creativity*, volume 82.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training. *Journal of.*

Ryan, J. 2018. *Curating Simulated Storyworlds*. Ph.D. Dissertation, University of California Santa Cruz, CA, USa.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Author Contributions

Gonzalo Méndez carried out the experiments with ChatGPT and wrote the initial draft of the paper. Pablo Gervás revised the draft and elaborated on previous work on neural models. Both authors revised the paper several times and jointly developed the discussion and the conclusions.

7. Demos

Steering latent audio models through interactive machine learning

Gabriel Vigliensoni[1] **and Rebecca Fiebrink**[2]
[1, 2]Creative Computing Institute, University of the Arts London, UK
[1]Centre for Interdisciplinary Research in Music Media and Technology, QC
g.vigliensoni@arts.ac.uk

Abstract

In this paper, we present a proof-of-concept mechanism for steering latent audio models through interactive machine learning. Our approach involves mapping the human-performance space to the high-dimensional, computer-generated latent space of a neural audio model by utilizing a regressive model learned from a set of demonstrative actions. By implementing this method in ideation, exploration, and sound and music performance we have observed its efficiency, flexibility, and immediacy of control over generative audio processes.

Introduction

Recent advances in neural audio synthesis have made it possible to generate audio signals in real time, enabling the use of applications in musical performance. However, exploring and playing with their high-dimensional spaces remains challenging, as the axes do not necessarily correlate to clear musical labels and may vary from model to model. In this paper, we investigate and propose a useful new approach based on interactive machine learning. This approach allows the performer to map the well-known, low-dimensional, human performance space to the high-dimensional generative audio model's latent space by providing training examples that pair the two spaces.

Background

Generative AI audio models

Generative AI audio models provide a data-driven approach to sound generation. These systems are designed to autonomously generate audio signals by learning from existing or custom datasets, capturing the underlying patterns and characteristics of the input data. However, historical systems for generative audio modelling and synthesis, such as WaveNet (Oord et al. 2016) and SampleRNN (Mehri et al. 2017)), have been challenging to integrate into creative environments due to their large computational complexity, poor signal quality, short temporal coherency, and lack of interaction means. Newer neural audio synthesis architectures and systems such as DDSP (Engel et al. 2020) and Jukebox (Dhariwal et al. 2020) have introduced advancements that addressed part of the previously mentioned issues. DDSP can model audio signals using small training datasets and can be steered in real time using pitch and amplitude as generative conditions, but only for monophonic instrument signals. Jukebox can generate a singing voice overlaid on top of complex, polyphonic music signal using text, genre, and artist labels as condition factors, but it requires massive computational power and datasets to be trained and lacks real-time control at generation time. The more recent architecture RAVE (Caillon and Esling 2021) addresses all the aforementioned issues in the context of modelling complex, polyphonic audio signals. However, given the potentially large dimensionality of the learned embedding and also the lack of labels for the latent space axes, there is a need to find a better way for real-time interaction and performing with such models.

Steering Generative AI

Real-time control in neural audio synthesis systems is important as it can enable performers to introduce the long-term temporal coherence often missing in these systems. That is, a generative model producing audio signals with short-term temporal coherence can still be used to generate longer structures if meaningful control is applied during generation. We next describe three main approaches to exerting control on the generative process.

Training data. In creative contexts, the choice of training dataset serves as the primary mechanism through which a human creator specifies what kind of content the machine should generate. This approach is often overlooked due to the extensive data and processing power required by most generative systems. However, working with small-scale datasets has been proposed as a means to allow greater human influence over generative AI systems in creative contexts, better aligning with creators' goals and ways of working (Vigliensoni, Perry, and Fiebrink 2022). In particular, when datasets are small, minor changes, such as the addition or removal of a few training examples, can significantly impact the trained model's behaviour.

Conditioning. In generative tasks, conditioning is a useful approach for controlling the generative process. By passing a certain condition to the network, the system can generate output conditioned on a specific variable. Conditioning can be applied when setting up the generative inference process (e.g., by using the artist or genre labels in Jukebox) or at

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

inference time (e.g., when conditioning DDSP with pitches and amplitude). In the case of RAVE, the generative process can be indirectly conditioned, such as by using sound content in a timbre transfer task. For instance, a beat track could serve as a MIDI-like clock, and the spectral content of an input signal can condition the output to generate a signal with similar frequency content.

Latent manipulation. This approach involves overriding latent dimension values with user input. For example, the RAVE architecture consists of an encoder that learns to map input audio data to a latent space and a decoder that learns to reconstruct the original data from the latent representation. When performing latent manipulation, one or more latent dimensions' values learned by the RAVE network can be replaced with the output from sliders controlled by a performer. Changes in values can be direct and absolute or relative to those generated by the encoder. In the latter case, arithmetic manipulation can be applied to the encoder output by adding a signal or multiplying it by a variable factor. From a performative perspective, latent manipulation is interesting because the performer can explore how the generative process changes when moving through orthogonal axes in the latent space. This exercise may help identify perceptual labels for specific dimensions. Alternatively, we propose below a novel approach to latent manipulation that uses supervised learning to map the human-performance space to the generative model's latent space.

Our Approach

The primary goal of this project is to devise and implement a real-time solution for steering a generative AI audio model towards a specific creative direction. Since the model has already undergone training, we cannot modify the underlying training data. Therefore, our sole means of interacting with the generative model involve conditioning it with specific features or performing latent manipulation. For example, we can condition the system by exciting the encoder with particular types of sounds, causing them to be projected into specific zones of the embedding and decoding similar sounds. Alternatively, we can perform latent manipulation by overriding the latent dimension values with user input.

The methodology we propose for performing and steering a neural audio model is inspired by research on machine listening systems. In this field, the most promising methods are hybrid systems that combine a data-driven approach informed by models of the perceptual and cognitive processes of the human auditory system (Heller et al. 2023). Similarly, our method to perform with a generative audio system involves utilizing a data-driven autonomous approach to learn the optimal representation for disentangling the audio data (e.g., using RAVE) and, subsequently, we work with the resulting embedding to identify creatively relevant or salient points within that space.

Interactive Machine Learning as a Mapping tool

Art- and music-making are non-teleological and purposeless activities in nature, not problems to be optimized (Audry 2021). As such, our approach to interacting with a neural audio model is centred on the curious and serendipitous exploration of its latent space. However, in order to facilitate more flexible and creative navigation of this space, we have explored the potential of interactive machine learning (IML) to map the human, well-known performance space onto the computer's label-less audio latent space.

IML (Fails and Olsen Jr 2003) is founded on the idea that training can be an incremental process in which the human and the machine collaborate to achieve a specific goal. In contrast to classical machine learning, where interaction with a model begins after it has been trained—usually following an extended offline period during which the algorithm iteratively optimizes to reach a certain model—IML as originally proposed by Fails and Olsen Jr involves a person iteratively experimenting with a machine learning model and tuning or steering its behaviour through changes to its training data. This human-machine interaction can happen over an extended period or in realtime, as the machine learns from human feedback and adjusts its models accordingly.

Such an IML approach has been used to create new gestural musical instruments since the introduction of the Wekinator tool (Fiebrink, Trueman, and Cook 2009), which enables people to iteratively construct and modify mappings from a human control space to sound synthesis parameters, through training examples pairing control-space coordinates and desired synthesis parameter values. To our knowledge, however, this approach has not previously been used to control generative models.

In this paper, we propose the IML paradigm as a tool for steering a generative audio model. The approach involves iteratively supplying training sets consisting of locations in the human-performance space paired with locations in the generative model latent space. We follow these steps:

1. We explore the latent space until identifying a point where an interesting zone emerges in terms of subjective creative possibilities. We describe this point with a descriptive auditory perceptual label. For example, "bright and loud" or "opaque and soft".
2. We select a source point in the performance space that should map to the target point in the latent space. Similar perceptual labels should be clustered together and can follow a meaningful progression in the performance space, such as arranging soft to loud sounds on the horizontal axis and from bright to opaque sounds on the vertical axis of a 2D controller.
3. We repeat the previous two steps as many times as desired, based on our creative aim. Ultimately, we will have a dataset comprising several pairs of source and target points linking the two spaces.
4. We instantiate the learning of a mapping between the performance and the latent space using the built dataset. A regression algorithm learns to map the points between the two spaces. These steps can be repeated to modify the mapping.

The mapping between the two spaces is shown in Fig. 1. The figure shows how a vector describing a point in "performance space" on the left is mapped onto a higher dimensional space given a series of models learned via regression.

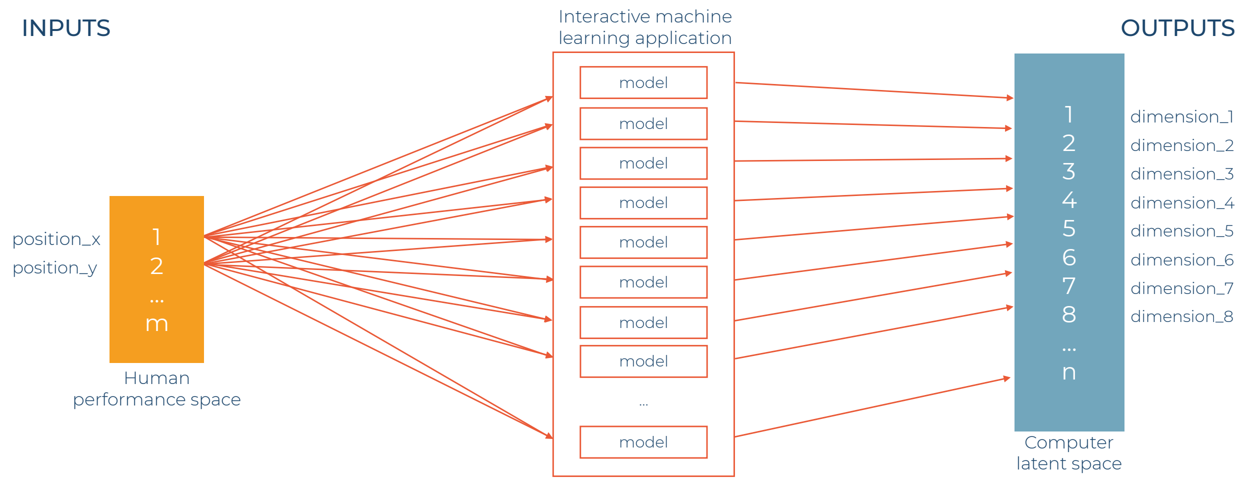


Figure 1: Interactive machine learning as a mapping tool. The low-dimension human-performance space is mapped to the high-dimensional computer-latent space. The mapping is done through a regressive task using a supervised learning approach.

For example, the input values could be the (x, y) coordinates when using a mouse on a canvas or an XY grid controller, or six values $(x_1, y_1, z_1, x_2, y_2, z_2)$ if using a six degrees of freedom controller such as a Gametrak.[1] We create one regression model per dimension of latent space, rather than one multi-dimensional model outputting a full latent space vector, to keep each modeling task simpler and thus trainable with fewer examples.

Experiments, Use, and Reflection

In our experiments, we have used the Wekinator and the FluCoMa (Tremblay, Roma, and Green 2021) toolboxes as frameworks for learning regression models using a supervised approach. Given that the number of training examples we have used is typically small (in the order of a few dozen), only a shallow (1 or 2 hidden layers) multilayer perceptron neural network is needed, facilitating very fast training and retraining. We have applied this method to map performance spaces where gestures are captured from on-screen and physical/gestural controllers using an arbitrary number of degrees of freedom (in our experiments, 2, 3, 6, and 15). These gestures have benn then mapped to steer RAVE latent audio models, encompassing a range from 4 to 64 latent dimensions.

In Figure 2, we present a graphical user interface of an instance of our approach using RAVE inside MaxMSP, and the FluCoMa package to learn a mapping between a human-performance space (a 2D mouse canvas in this case) to the computer-latent space (8D in this case). Once a mapping is learned, the selected zones of the latent space are mapped onto the performance space, and the performer plays the performance space.

Some of this experimentation has taken place in ideation and live performance contexts as part of the first author's preparation for Visiones Sonoras 18,[2] beat-based electronic music performances, both solo and in a duo with sound artist dedosmuertos, in which IML-generated models were employed for real-time gestural control of RAVE. We have also tried this setup in DJ sessions where the digital turntable's output signal has been timbre-transferred using audio models and our IML-enabled manipulation of the latent space.

Our approach has allowed us to interact and play with latent audio models in a straightforward and flexible way. In particular, it has enabled us to move between distant points in the latent space efficiently and relatably in the human performance space. Given the small amount of training data needed to learn the mappings, we have even retrained the system during the performance. The mappings between the spaces are not discrete but continuous, resulting in additional control as we can engage in constant subtle modulation of the latent space decoding, leading to continuously changing audio signals. In our experiments, we have experienced the immediacy of our approach to control over the generative audio process.

The most significant drawbacks we have experienced in performance are the latency of the generative system, which introduces a delay between the human gesture and the resulting action, and the potential existence of problematic zones in the latent space that can lead to unexpected loud sounds. While the former issue is inherent to digital audio buffering, we have addressed the latter by employing heavy limiting, rehearsing, and familiarizing ourselves with the spaces.

A video demonstrating the training of a mapping model and its use in performing with a high-dimensional audio latent space using a mouse and a Gametrak controller can be accessed at `https://bit.ly/iccc2023`.

Some key insights from this experimentation include: (i) Using shallow neural networks such as those in Wekinator and FluCoMa was adequate for building useful mapping functions that usually matched our intention. (ii) Even minimal training data (e.g. a few dozen examples) usually suffices to create a useful and playable mapping between the two spaces. The small size enables the training and retrain-

[1] https://en.wikipedia.org/wiki/Gametrak

[2] https://en.cmmas.com/vs18

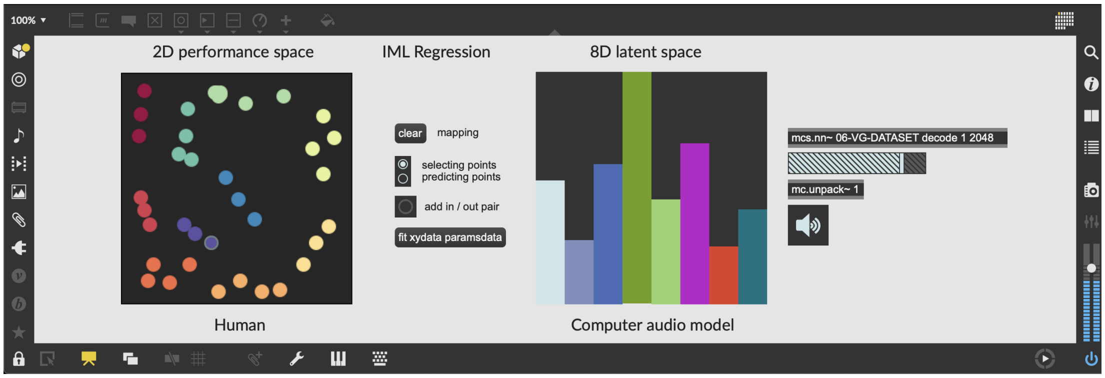


Figure 2: Graphical user interface showing the mapping between the human-performance space to a higher-dimensional latent audio model. In this example, an 8-dimensional space is controlled by means of a 2D space.

ing of models even at performance time. (iii) Sometimes, due to the small amount of data, our method yields models that do not perfectly match intentions. However, these individually crafted models of interaction can still prove to be useful and inspiring in a creative context. (iv) This approach facilitates creation of control trajectories that allow for drastic or smooth transitions between points in the latent space. (v) The IML approach to mappings promotes fast prototyping, flexibility in mappings creation, and immediacy of control. (vi) In performance, this approach allows us to overcome the issue of short temporal coherence often found in generative neural audio systems. Because a performer has control over the generative process, they can maintain a longer window of coherence and manipulate sound and music motives and tension more effectively. This can be achieved, for instance, by revisiting or introducing new timbres or motifs.

Conclusion

We have described how IML can enable performers to map from real-time control vectors—from on-screen or physical controls—to creatively relevant or salient points within a latent space. We have found that IML can be an effective tool for enabling real-time, performative interactions with generative models, even when the latent space of a model is high-dimensional and its dimensions do not neatly correspond to perceptual attributes.

Acknowledgments

This research has been supported by Fonds de recherche du Québec – Société et culture (FRQSC) through a research creation grant (Ref. 2022-B5R-310633). Important parts of this work used Digital Research Alliance of Canada's High Performance Computing resources.

References

Audry, S. 2021. *Art in the Age of Machine Learning*. Cambridge, Massachusetts: The MIT Press.

Caillon, A., and Esling, P. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv:2111.05011.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. arXiv:2005.00341.

Engel, J.; Hantrakul, L.; Gu, C.; and Roberts, A. 2020. DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations (ICLR 2020)*.

Fails, J. A., and Olsen Jr, D. R. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, 39–45.

Fiebrink, R.; Trueman, D.; and Cook, P. R. 2009. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of the 9th International Conference on New Interfaces for Musical Expression*, 280–285.

Heller, L. M.; Elizalde, B.; Raj, B.; and Deshmukh, S. 2023. Synergy between human and machine approaches to sound/scene recognition and processing: An overview of ICASSP special session. In *Special session on "Synergy between human and machine approaches to sound/scene recognition and processing" at the 2023 ICASSP*. arXiv:2302.09719.

Mehri, S.; Kumar, K.; Gulrajani, I.; Kumar, R.; Jain, S.; Sotelo, J.; Courville, A.; and Bengio, Y. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. In *International Conference on Learning Representations (ICLR 2017)*.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. WaveNet: A generative model for raw audio. arXiv:1609.03499.

Tremblay, P. A.; Roma, G.; and Green, O. 2021. Enabling programmatic data mining as musicking: The Fluid Corpus Manipulation toolkit. *Computer Music Journal* 45(2):9–23.

Vigliensoni, G.; Perry, P.; and Fiebrink, R. 2022. A small-data mindset for generative AI creative work. In *Proceedings of the Generative AI and Computer Human Interaction Workshop (GenAICHI, CHI 2022 Workshop)*.

Calliffusion: Chinese Calligraphy Generation and Style Transfer with Diffusion Modeling

Qisheng Liao,[1] Gus Xia,[1,2] Zhinuo Wang,[2]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]New York University Shanghai
qisheng.liao@mbzuai.ac.ae, gus.xia@mbuzai.ac.ae, zw2375@nyu.edu

Abstract

In this paper, we propose Calliffusion, a system for generating high-quality Chinese calligraphy using diffusion models. Our model architecture is based on DDPM (Denoising Diffusion Probabilistic Models), and it is capable of generating common characters in five different scripts and mimicking the styles of famous calligraphers. Experiments demonstrate that our model can generate calligraphy that is difficult to distinguish from real artworks and that our controls for characters, scripts, and styles are effective. Moreover, we demonstrate one-shot transfer learning, using LoRA (Low-Rank Adaptation) to transfer Chinese calligraphy art styles to unseen characters and even out-of-domain symbols such as English letters and digits.

Introduction

Chinese calligraphy, which is the artistic writing of Chinese characters and a prominent form of East Asian calligraphy, can be seen as a distinctive form of visual art. There are five Chinese calligraphy scripts, regular (楷), semi-cursive (行), cursive (草), clerical (隶), and seal (篆) script. Regular script is the most common script for writing nowadays. Semi-cursive script is faster to write compared with regular script but is still easily readable. Cursive script is known for its speedy writing style, but it can be challenging to read. Clerical script and seal script nowadays are mainly used for artistic purposes. Besides, each famous calligrapher has his or her own style. Even when they write the same character in the same script, the calligraphy may look very different. For example, Figure 1 shows 10 samples of the same character. Here, each column belongs to a different script, and for each script, we show two samples of different styles.

Recently, we see a trend in generating Chinese calligraphy using AI, including Zi2zi(Tian 2017), CalliGAN(Wu, Yang, and Hsu 2020), and ZiGAN(Wen et al. 2021). Most of them adopt a GAN (Generative Adversarial Network)(Goodfellow et al. 2014) architecture, training on paired data of printed font and handwritten font while performing image-to-image translation during inference. Despite some effort on improving the data efficiency (Zhou et al. 2021), two major challenge remains: (1) to generate *high-quality* calligraphy, and (2) to apply effective *controls* on characters, scripts, and calligraphers' styles.

Figure 1: 10 samples of the character ”风” (wind).

If we visualize the calligraphy artworks in the 3-D space of characters, scripts, and styles, the distribution of the samples would be very *sparse*. There are thousands of Chinese characters, while most calligraphers' work collections only cover a small portion of the characters in particular scripts. In this paper, we aim to model the sample distribution in the 3-D space and generate calligraphy of any character, script, and style.

To this end, we introduce a new method for generating Chinese calligraphy with Denoising Diffusion Probabilistic Models (DDPMs)(Ho, Jain, and Abbeel 2020). In particular, we control the model using external conditions based on Chinese text descriptions of character, script, and style. During training, we utilized labeled calligraphy images, while during inference, we used description texts to control the generation process. Notably, unlike most previous studies that rely on GANs and require an input image for generation, our method does not necessitate the use of any images during inference.

Besides with-in distribution generalization, we also utilize Low-Rank Adaptation of Large Language Models (LoRA)(Hu et al. 2021) to achieve out-of-distribution *style transfer* via one-shot fine-tuning. Experiments show that such an approach can transfer existing scripts and styles to unseen characters and even out-of-domain symbols such as English letters and digits.

Our model could be very useful for individuals engaged in the process of learning Chinese calligraphy. A common approach to learning Chinese calligraphy is to study the art-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

work of famous calligraphers and imitate their styles. However, the artwork resources of specific calligraphers are usually limited, and it is almost impossible to obtain any character for a specific calligrapher. With the aid of our model, learners can overcome such limitations, generating artworks of any character, any script, and any style.

In summary, the major contributions of our Calliffusion system are:

- As far as we know, it is the first diffusion model for generating high-quality Chinese calligraphy artwork.
- The controllable generation is effective. The conditional model can generate calligraphy in any character, script, and calligrapher's style.
- The style transfer technique is effective. Our one-shot fine-tuning technique can adapt certain scripts and writing styles to unseen Chinese characters and even English letters and digits.

Methods

Diffusion Model

In our research, we used a U-net(Ronneberger, Fischer, and Brox 2015) model as the backbone model and used DDPMs sampling, which include a forward process (diffusion) that progressively disturbs the organization of the data x_0, and a reverse process (denoising) that is trained to restore the initial data x_0 from the corrupted input. In this context, x_0 refers to the calligraphy image. The forward process involves the addition of Gaussian noise in N diffusion steps as shown in equations 1 and 2 below.

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{2}$$

The variance scheduling parameters $\beta_1, \beta_2, \ldots, \beta_N$ are employed to regulate the diffusion process. On the other hand, the reverse process requires the model to define a Markov chain that sequentially rebuilds the calligraphy image x_0 from a disturbed input x_N, which follows a normal distribution $N(0, I)$. The equations 3 and 4 below show the reverse process.

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{3}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{4}$$

While in the process of training, we aim to minimize the target by optimizing the model parameters represented by ϵ_θ as equation 5, where t is uniformly sampled from $[1, N]$ and $\epsilon \sim \mathcal{N}(0, I)$, $a_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

$$L(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right] \tag{5}$$

Adding Controls with External Conditions

In order to control the generations, we rely on three conditions, i.e., characters, scripts, and styles. In specific, We use a short description of Chinese text input, such as '人字 隶书 曹全碑' ('Ren Character, Clerical script, Caoquanbei') to control the generations. The text consists of three parts, and a space separates each part. The first part of the text determines the character, the second part controls the script, and the last part determines the calligrapher's style. The input text is then passed through a pre-trained Chinese BERT model (Devlin et al. 2018) to obtain cross-attention embeddings. These embeddings are combined with the image during the training of the diffusion model. The structure of this conditional model is illustrated in Figure 2.

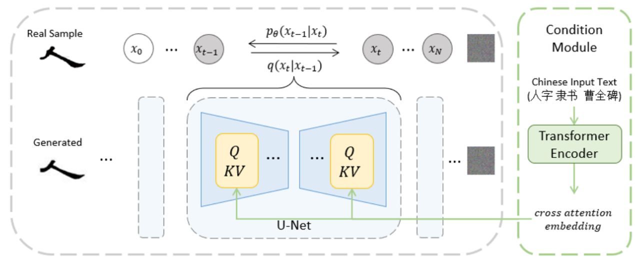


Figure 2: The diffusion model structure with one cross-attention condition that comes from a Transformer encoder.

Style Transfer via Fine-tuning

Based on the conditional diffusion model, our Calliffusion system can further transfer the scripts and styles to unseen characters and out-of-domain symbols via one-shot fine-tuning. During the fine-tuning process, we only need to provide the model with a single image of the new character or symbol, either letting its script or style be specified or not specified. After that, the system can generate new calligraphy by applying a script and a style to *that* character or symbol.

The fine-tuning technique is based on LoRA, which is a training technique that speeds up the training process of large models while reducing memory consumption. LoRA achieves this by adding update matrices, which are rank-decomposed weight matrices, to the existing weights, and only updating the newly added weights during training. By keeping the previously pretrained weights frozen, the model is protected against catastrophic forgetting, where it loses previously learned information during further training. Additionally, the rank-decomposition matrices used in LoRA have significantly fewer parameters compared to the original model, making the trained LoRA weights easily transferable and portable.

Training

Dataset

We collected our own dataset by downloading copyright-free Chinese calligraphy images from online. The dataset

(a) 5 generated character 'Dong' in different scripts.

(b) 5 generated character 'Jin' in different scripts.

(c) 4 characters in clerical script and Caoquanbei's (曹全碑) style. The first image is generated by our model and the real sample for this character in Caoquanbei's style does not exist. The other three images are real Caoquanbei's calligraphy.

(d) 4 characters in semi-cursive script and Wang Xizhi's (王羲之) style. The first image is generated by our model and the real sample for this character in Wang Xizhi's style does not exist. The other three images are real Wang Xizhi's calligraphy.

(e) 4 characters in cursive script and Mao Zedong's (毛泽东) style. The first image is generated by our model and the real sample for this character in Mao Zedong's style does not exist. The other three images are real Mao Zedong's calligraphy.

(f) 4 characters in seal script and Wang Kuaijishike's (会稽石刻) style. The first image is generated by our model and the real sample for this character in Wang Kuaijishike's style does not exist. The other three images are real Kuaijishike's calligraphy.

(g) Generated calligraphy for a sentence of a poem. The conditions are regular script and Yan Zhenqin(颜真卿).

(h) Generated calligraphy for a sentence of a poem. The conditions are semi-cursive script and Su Shi(苏轼).

(i) Generations based on with one-shot fine-tuning. The conditions for generation are different from the conditions in fine-tuning but the generated calligraphy images have the features of those conditions.

Figure 3: Qualitative results of our Calliffusion system.

includes images from 5 scripts, featuring 3975 unique characters and 1431 artists. During the preprocessing stage, we applied a threshold and only retained characters with more than 10 samples, resulting in a reduced dataset of 2025 characters and 1387 artists.

Additionally, for style transfer with English letters and numbers, we utilized a handwriting dataset from a previous study (De Campos et al. 2009).

Hyperparameters

We utilized the "diffusers" package(von Platen et al. 2022) in Python as the underlying framework for our diffusion models. We configured four blocks in the U-Net architecture with dimensions of 320, 640, 1280, and 1280, each consisting of two layers. We used a Chinese BERT(Devlin et al. 2018) model to obtain cross-attention embeddings with a size of 768 from the input text. The sample size was set to 64, and the batch size was set to 16. We employed the Adam optimizer with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-6}. The training was conducted on two NVIDIA A100 40G GPUs for a total of 120 hours.

Calligraphy Generation Examples

Style-free Generation

Though we used three different conditions to train our models, we do not have to specify all of them during generation.

	Regular		Semi-cursive		Cursive		Clerical		Seal		Total	
	Script	Character	Script	Character	Script	Character	Script	Character	Script	Character	Script	Character
Real Samples	0.91	0.93	0.83	0.81	0.88	0.68	0.96	0.83	0.97	0.81	0.88	0.78
Generated w/o style condition	0.92	**0.94**	**0.86**	0.89	**0.89**	**0.72**	0.94	0.87	0.97	**0.80**	0.91	0.84
Generated w/ style condition	**0.96**	**0.94**	**0.86**	**0.95**	0.88	0.64	**0.97**	**0.91**	**0.99**	0.79	**0.93**	**0.85**

Table 1: The performance of our generated data in different scripts in accuracy

In this section, we show some style-free generation examples by only conditioning on scripts and characters during inference time. Figure 3(a) and Figure 3(b) show generated artworks for two characters, each rendered with 5 different scripts. We see that our models are capable of producing high-quality Chinese calligraphy images, and the controls applied to both character and script are effective.

Stylistic Generation

We randomly choose 4 characters and render them in an "unfamiliar" style, in the sense that the character-style pair never appears in our dataset. In Figure 3(c) to Figure 3(f), each rendered example is listed together with several real artworks in the corresponding style to showcase the style similarity and consistency. Here, in each of these subfigures, the first character was generated by our model and the other three are real samples. These examples demonstrate that the control on style is effective, and our later subjective evaluation reveals that even people who know these styles well have difficulty spotting real and generated artworks.

Transfer learning with One-shot Fine-tuning

During our training, we intentionally leave out a common Chinese character '明' (bright). Later, we handpick two samples, one in regular script by Liu Gongquan and the other in clerical script by Tangtaizong, to fine-tune the model, respectively. After fine-tuning, our model acquired the knowledge of this character and can apply it to other calligraphers' styles. As depicted in the left side of Figure 3(i), we generate '明' in regular script with Yan Zhenqin and in clerical script with Caoquanbei.

For digits and English letters, which are certainly not included in our dataset, we pick '4' and 'a' to conduct fine-tuning. Even with just one-shot, we do not set any specific script or style conditions but only inform the model that the characters are '4' and 'a'. During inference, we incorporate seal script as a condition, and the resulting generated images, as shown on the right-hand side of Figure 3(i), exhibit the features of seal script.

Experimental Results

Objective Evaluation

We used an off-the-shelf pre-trained classifier to recognize the generated images. The classifier is a multitask classifier., whose backbone model is a Res-Net model with two classification embedding layers, one for scripts and one for characters. The generated corpus consists of 2000 images. The first 1000 images are generated by conditioning on the 200 most common characters, each with 5 scripts. We also keep the setting but select 5 famous calligraphers' styles as an extra condition to generate another 1000 images.

The results presented in Table 1 show that our generated calligraphy is highly similar to real calligraphy. Our generated samples have slightly higher accuracy than the test data (which are real artworks) of the pre-trained classifier. Furthermore, adding style conditions marginally improve the overall accuracies.

Subjective Evaluation

We designed a survey with three types of questions:

Identify the fake artwork: For each question, we randomly choose a calligrapher's style and select one generated sample produced by our model. The character of the generated sample has never appeared in the collection of the calligrapher's work. Then, we list the generated sample with three genuine artworks composed by the same calligrapher (similar to the layout shown in Figure 3(c) to Figure 3(f)), asking subjects to identify which of the four choices is generated by an AI model.

Identify the real artwork: The setup is similar to the first type, but the task is to tell the real artwork from the fake ones generated by the model.

Identify the script after transfer learning: For each question of the third type, we use either an English letter or digit generated by our fine-tuned transfer-learning model conditioned with specific scripts and ask subjects whether they can point out the scripts of generated characters. (The generated results are similar to the samples in Figure 3(I).)

The survey comprises a total of 10 questions. The first two types of questions contain 4 options each, and a lower accuracy indicates that our generated Chinese calligraphy is highly similar to genuine calligraphy. The third type of question presents 5 options, and a higher accuracy indicates that our generated calligraphy for non-Chinese characters exhibits the characteristics of Chinese calligraphy scripts, making it recognizable to subjects.

Table 2 presents the average accuracy and p-value of z score hypothesis testing for each type of question. We collected responses from 150 individuals in China, out of which 87 claimed to have practiced Chinese calligraphy or know the scripts and style used in the survey. The null hypothesis in this study is that the accuracy for each question is equal to random guessing (25% for questions with 4 options and 20% for questions with 5 options).

For the first two types of questions, the accuracy for individuals with previous knowledge of Chinese calligraphy is slightly higher than random guessing, whereas, for those who are unfamiliar with calligraphy, the accuracy is slightly lower. The p-values show that the results are not significantly different from random guess. In contrast, the third

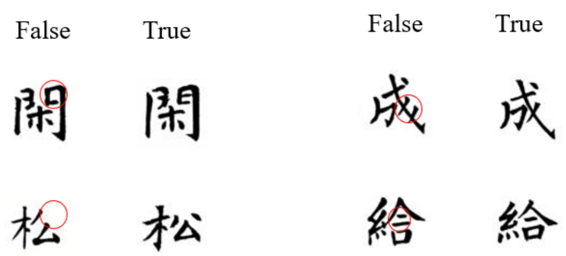
	Know Calli	Don't know	Total
No.	87	63	150
Q.	Acc(P-Val)	Acc(P-Val)	Acc(P-Val)
1↓	0.275(0.296)	0.245(0.853)	0.263(0.459)
2↓	0.286(0.293)	0.246(0.917)	0.269(0.458)
3↑	0.796(***)	0.579(***)	0.706(***)

Table 2: The accuracy and p-value of each type of question in our survey.

type of question revealed that around 70% of the subjects were able to identify the calligraphy script characteristics in our generated non-Chinese symbols, and the p-value indicates that this result is significant at $p < 0.001$.

Limitation

In this section, we presented examples of unsuccessful generated outcomes that could potentially pass an objective classification assessment but can be easily identified by humans familiar with the Chinese language. These failures can be categorized into two primary types: the missing of certain strokes, shown in Figure 4(a), or the addition of unnecessary strokes, shown in Figure 4(b). According to our experiment, we discovered that increasing the amount of training data and conducting more training epochs can lead to a reduction in the number of generated failures.



(a) Generated failures with missing strokes.

(b) Generated failures with unnecessary extra strokes.

Figure 4: Comparison of unsuccessful and successful results of our Calliffusion system.

Conclusion

In this paper, we introduce a conditional diffusion model for generating Chinese calligraphy. We demonstrate that our model can produce high-quality calligraphy by conditioning it with various combinations of characters, scripts, and styles. Additionally, we can generate previously unseen Chinese characters or even non-Chinese symbols using a one-shot transfer learning with LoRA. The artworks produced by our model undergo assessment through both objective and subjective evaluations. The objective evaluation demonstrates that our generated calligraphy exhibits exceptional accuracy when classified by a pre-trained classifier. The subjective evaluation indicates that when our generated samples are compared with authentic calligraphy, discerning between the two becomes exceedingly challenging for human observers.

Moving forward, our aim is to delve into the realm of few-shot style transfer learning for novel styles and scripts. Currently, we have the capability to perform style transfer for new characters or symbols using a one-shot approach. The future plan is to discover an effective method to learn new scripts beyond the five conventional Chinese calligraphy scripts or acquire new styles from the handwriting of any individual, which could make our model become even more valuable and versatile.

References

De Campos, T. E.; Babu, B. R.; Varma, M.; et al. 2009. Character recognition in natural images. *VISAPP (2)* 7(2).

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33:6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Tian, Y. 2017. zi2zi: Master chinese calligraphy with conditional adversarial networks.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models.

Wen, Q.; Li, S.; Han, B.; and Yuan, Y. 2021. Zigan: Fine-grained chinese calligraphy font generation via a few-shot style transfer approach. In *Proceedings of the 29th ACM International Conference on Multimedia*, 621–629.

Wu, S.-J.; Yang, C.-Y.; and Hsu, J. Y.-j. 2020. Calligan: Style and structure-aware chinese calligraphy character generator. *arXiv preprint arXiv:2005.12500*.

Zhou, P.; Zhao, Z.; Zhang, K.; Li, C.; and Wang, C. 2021. An end-to-end model for chinese calligraphy generation. *Multimedia Tools and Applications* 80:6737–6754.

Artist Discovery with Stable Evolution

Simon Colton,[1] **Amy Smith,**[1] **Blanca Pérez Ferrer**[2] and **Sebastian Berns**[1]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]Etopia Center for Arts and Technology, Zaragoza, Spain
s.colton@qmul.ac.uk

Abstract

We describe the Stable Evolusion system, with which users can evolve text prompts for use in the Stable Diffusion text-to-image generator, and view the resulting imagery. The system is designed for simplicity, enabling users to explore a space of styles visualising content of their choosing. We automate elements of the process with both a semantic search and a novelty-based search, to direct the user towards images of interest, and to help maintain diversity respectively. In addition, by combining the approach with Google Lens image searching, we enable the discovery of human artists and their artwork via the pre-generation of images similar to theirs.

Introduction and Background

A new way of producing high-quality generative art has recently emerged, namely the employment of neural text-to-image generators such as MidJourney (midjourney.com), DALL-E (openai.com/product/dall-e-2) and Stable Diffusion (dreamstudio.ai). The reaction to this from artists has been mixed, with some rightly upset that their artwork has been used to train generative deep learning models without permission, possibly infringing copyrights. While finding the right prompt for the generators is often not easy, the ease of use of these systems is perceived to threaten the livelihoods of commercial artists. To make matters worse, the image generators are able to fairly faithfully reproduce certain artists' styles, if prompted with their name, which could further effect livelihoods and legacies.

Other artists have embraced the new creative affordances that have arisen. For instance, photographer Boris Eldagsen recently won a category in the Sony World Photography competition with an AI-generated image, which caused some controversy (Williams 2023). To Eldagsen, the new way to generate images: "... is setting me free ... the boundaries I had in the past – material boundaries, budgets – no longer matter". He points out that the art of choosing the right prompt is not as easy as critics such as (McCormack et al. 2023) suggest. Indeed, dozens of websites where prompts can be downloaded, exchanged or purchased have sprung up. Eldagsen further points out that:

> "for the first time in history, the older generation has an advantage, as AI is a knowledge accelerator. Two thirds of the prompts are only good if you have knowledge and skills, when you know how photography works, when you know art history." (Williams 2023)

We describe here the *Stable Evolusion* system which helps novice users of the Stable Diffusion image generator to produce images via the evolution of prompts. This addresses somewhat the difficulty people have in writing prompts to achieve imagery of their liking, as the system supplies terminology from art practice and history without the user needing to know these. While not difficult to write a prompt for text-to-image generation, it is difficult to write the right prompt to achieve the kind of imagery required for a particular project. We apply Stable Evolusion to the discovery of human-produced artworks via Google image search, which could in a small way benefit commercial artists, balancing somewhat the difficulties they've encountered recently.

Stable Evolusion is written in a Colab notebook (Bisong 2019), built on top of the following technologies:

- **Stable Diffusion**. Released by Stability AI, this is a text-to-image generation system employing a *latent diffusion* model (Rombach et al. 2021) which iteratively de-noises a Gaussian noise image conditioned with a text prompt.
- **CLIP**. Released by OpenAI, this comprises two models which can encode text and images respectively into the same latent space (Radford et al. 2021). As described in (Colton et al. 2021), CLIP can be used to calculate semantic similarities between images and texts.
- **Vendi**. This is a method for estimating the diversity in a set of media such as images (Friedman and Dieng 2022). It can employ any similarity function, such the distance between image embeddings in a latent space. We use the inception machine vision model for this (Szegedy et al. 2016).
- **Google Lens**. This is a suite of image recognition systems which can be employed to search the internet for images similar to a given image (Conditt 2017).

System Description

The Stable Evolusion system has two ways to employ Stable Diffusion. Firstly, users can supply a key for the DreamStudio API offered by Stability AI (dreamstudio.ai), hence the notebook runs on a CPU, with image generation in the cloud. Alternatively, the notebook has code from the *HuggingFace Diffusers* package (Patil et al. 2022), so image generation can be performed on a Colab-supplied GPU. Generation of a 512x512 pixel image with the API takes around 5 seconds per image (network dependent); a standard Colab GPU takes around 11 seconds; a premium GPU takes around 3 seconds. For the experiments here, we used the Diffusers package.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

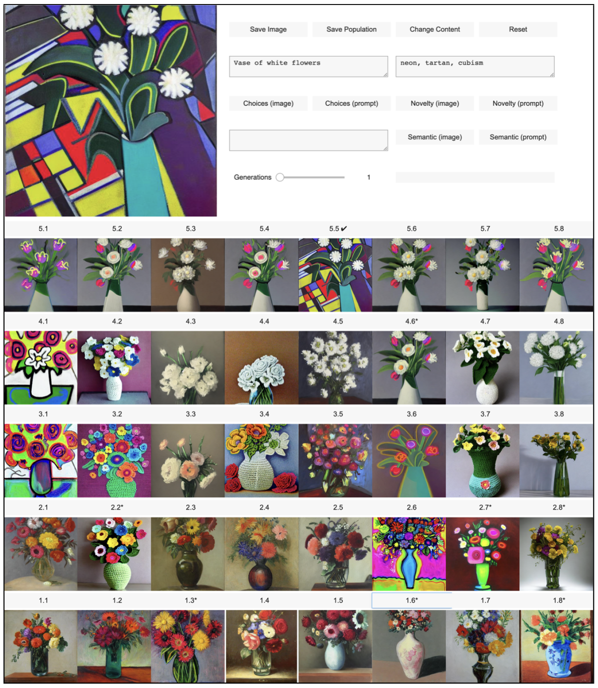


Figure 1: Screenshot from the Stable Evolusion notebook.

Stable Evolusion employs a straightforward evolutionary approach where a genome consists of:

- A random **seed** as an integer between 1 and 100,000
- A **content text** describing the scene to be depicted
- A list of visual **modifiers** which are short text phrases

At the start of a session, the user supplies an initial content text such as "painting of a chair", and then starts the process. The system produces a first generation of 8 images using the unmodified content text as a prompt to Stable Diffusion and presents these to the user. The user can then select any image(s) they want to develop and produce a new generation from; if they choose none then it is assumed the next generation should be produced from all of the current generation. Users are free to select images from any previous generation with selections wiped after they have been used. Five generations are shown in the user interface, as per figure 1.

We have experimented with two evolutionary mechanisms for producing novel prompts, namely *extension* and *mutation*. The extension process simply adds a modifier to the list for a selected genome. The prompt for generating an image is produced by concatenating the modifiers to the content. Users can decide whether to keep the random seeds of their chosen genomes when generating the next round. As we see in experimentation below, child images generated with the same seed, but slightly different modifiers are substantially more similar to their parent than those where the seed is changed. Hence, keeping seeds enables users to make smaller steps in the possibility space, while changing seeds affords bigger steps. Buttons in the UI marked *(image)* direct Stable Evolusion to keep seeds, those with *(prompt)* direct it to change them. When a new generation is made, the older generations move down in the GUI to make space.

The pre-selected modifiers have been hand-curated from visual art history and practice, with adjectives and short phrases describing: media (e.g., *oil painting*); movements (e.g., *impressionism*); styles (e.g., *unfinished*); moods (e.g., *melancholic*); colours (e.g., *vibrant colors*); patterns (e.g., *tartan*); textures (e.g., *denim*); lighting (e.g., *moonlight*); and materials (e.g., *plastic*). After 25 such modifiers have been added to a genome, the entire prompt usually comprises more than the 77 tokens which can be accepted by Stable Diffusion, so modifiers 26, 27, etc., will not change the image. Hence at this stage, instead of extending the modifiers, one of them is changed (mutated) to a different one. As we see in the experiments below, this produces similar results to extending modifiers. If desired, the user can change the text prompt during a session, in which case, the system produces 8 new images which have the same seed and modifiers as in the most recent generation, but with the new content text. If the content text is related to the previous one, new images usually look quite similar to those in the row below it, which we found gives a satisfying level of control.

In the session depicted in figure 1, the user chose the prompt "Vase of flowers". In the first generation (at the bottom of the screenshot), 8 images were generated from just this prompt. The user chose images 1.3, 1.6 and 1.8 from these (marked with an asterix) by clicking on the buttons above them. In the next generation, single modifiers were added to the prompt and 8 more images were generated. The user chose images 2.2, 2.7 and 2.8, and the modifiers for these were extended into 8 images for generation three. For generation four, the user changed the prompt to "Vase of white flowers" and we can see that the images look somewhat like those directly below them, but with (more) white flowers. The user chose a single image (4.6) for the fifth generation and produced 8 variations of this by clicking on the 'Choice (Image)' button to retain the seed while extending the modifiers in 8 different ways. The large image showing is 5.5, with the modifiers *neon*, *tartan* and *cubism*.

To increase ease of use, the bottom part of the GUI enables users to force a search over multiple generations. These are guided by one of the following processes, which are each iterated for a user-given number of generations:

- **Novelty-based search**. Here, Stable Evolusion chooses to evolve the three most novel images in the current generation and extends/mutates their modifiers into the next generation. Novelty for an image, I, is determined by how much the Vendi diversity score reduces when calculated before and after I is removed. We experiment with alternatives below.

- **Semantic search**. Here, the user can supply a secondary *target phrase*, T, which is used in selecting images for evolution. In particular, the CLIP similarity to T for each of the images in the current generation is calculated, and the three most similar are evolved in the next round. An example of how this drives the search is given in figure 2, showing five generations of chair images progressing towards the target phrase 'psychedelic'. Naturally, the user could instead supply content text including the target phrase, but this can quite drastically change the images, and it is often preferable to evolve towards the target using a semantic search.

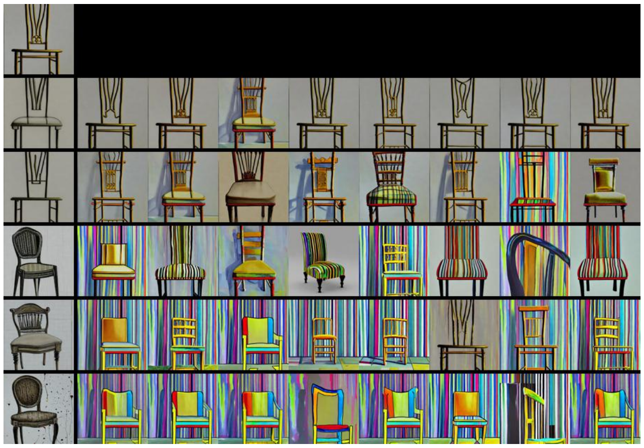


Figure 2: Images from a session automatically evolving chair images with the target word 'psychedelic'.

Experiments and Results

To study the effects of the extension and mutation evolutionary operators, we generated images for five content texts, namely 'downtown manhattan' (shortened here to city); 'painting of a chair' (chair); 'seascape with a boat' (seascape); 'vase of flowers' (flowers) and 'modern architecture building' (building). For each content text, T, we generated a parent image from a random genome with n modifiers, where n ranges from 0 to 24. We then extended these into a child genome by adding a single modifier, then produced a child image. For each (T, n), we produced parent/child pairs both keeping and changing the seed in the child, repeated for 5 different random genomes. This produced 1,250 parent/child pairs, over which we calculated the CLIP similarity and plotted relevant averages in figure 3(a).

We undertook a similar experiment with mutation rather than extension evolving the parent genome into the child, again with results plotted on figure 3(a). As expected, we see that when the child images are produced using the same seed as the parent, the CLIP similarity is substantially higher than when the seed is changed. We expected the similarity of parent/child pairs to increase in line with the position of the altered/added modifier, as later words in a prompt have lower effect on images in general. This trend is certainly observed when the child shares its parent's seed, with the similarity raising from around 0.88 to around 0.94. However, when changing the seed, this appears to be sufficiently disruptive to images that this trend is not observed.

	Keep seed			Change seed		
Content	Intra	Inter	Parent	Intra	Inter	Parent
City	0.833	0.659	0.916	0.789	0.663	0.826
Chair	0.864	0.678	0.933	0.828	0.685	0.842
Seascape	0.851	0.654	0.923	0.800	0.658	0.826
Flowers	0.847	0.647	0.922	0.805	0.659	0.838
Building	0.839	0.661	0.910	0.782	0.659	0.810
Average	0.847	0.660	0.921	0.801	0.665	0.828

Table 1: Average CLIP similarities over pairs of images.

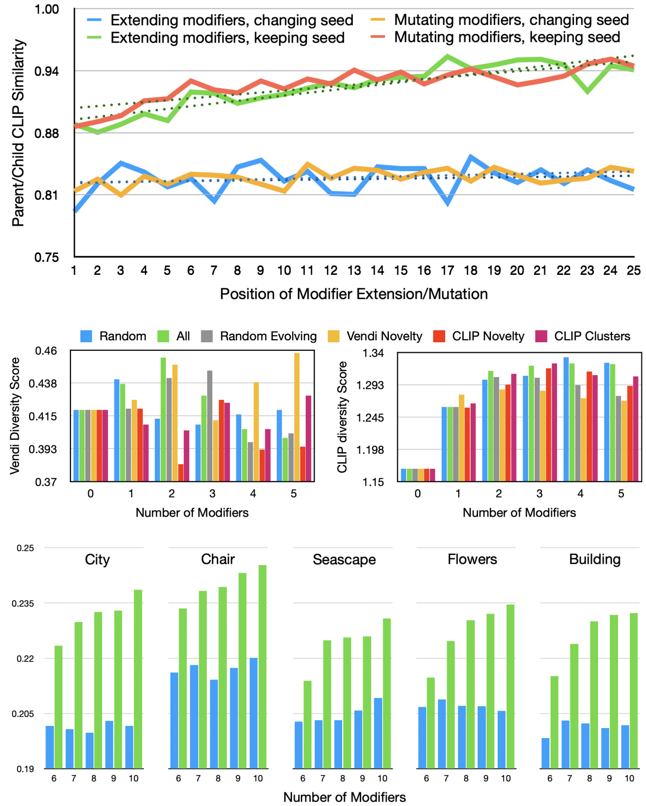


Figure 3: (a) Child/Parent image CLIP similarity, as the number of modifiers increases from 1 to 25, and the position of the mutated modifier increases. (b) Vendi and CLIP diversity measures over the different selection methods. (c) Comparison of CLIP similarities to target texts for semantic (green bars) and random (blue bars) searches.

For further context about CLIP similarities, in table 1, we record the average CLIP similarities over all pairs of images with (a) the same content text [intra] (b) different content texts [inter] and (c) a [parent]/child relationship. We see that CLIP similarities go as low as 0.65 inter categories, which highlights that relatively high similarities of up to 0.94 seen in figure 3(a) indicate strong visual similarity between parent and child images. This is borne out under visual inspection: often when changing the 25th modifier in a genome and keeping the seed for the child, it looks very similar indeed to its parent. Finally, we note that the graphs in figure 3(a) for extension and mutation are roughly similar. Hence there shouldn't be a noticable difference when mutation takes over. In practice, when mutations occur after 25 modifiers are added, they are applied randomly to a modifier in the final 5, and we've found this provides good continuity.

A reasonable use case for the novelty search is when a user starts a session and wants a diverse set of images to choose from initially, produced over, say, five generations. We experimented with the following six different mechanisms for producing 5 generations at the start of a session:

- **Random:** each genome is generated randomly with n modifiers for generation number n.

- **All:** all genomes in the current generation are evolved.
- **Random Evolving:** three genomes are selected randomly for evolution.
- **Vendi Novelty:** genomes for the three images which reduce the vendi diversity score the most are selected.
- **CLIP Novelty:** genomes for the three images with least total similarity to the other images are selected.
- **CLIP Clustering:** each genome is given a 7-entry vector *profile* by calculating the CLIP similarity between its image and the other images. This profile is used in a K-means clustering process to produce 3 clusters, from each of which a genome is selected randomly for evolution.

For each of the content texts above, over 5 trials each, an initial 8 images were generated from modifier-free genomes. The 8 genomes were then evolved over five generations via extending the modifiers (changing the seed), in six seperate sessions, i.e., one for each of the above generation mechanism. The diversity of each generation was estimated in two different ways: (a) using the Vendi diversity score, which essentially calculates the exponential of the Shannon entropy of a similarity matrix's eigenvalues (Friedman and Dieng 2022), with similarity being the cosine distance between embeddings of images in the inception latent space (Szegedy et al. 2016), and (b) the reciprocal of the average CLIP similarity over every pair of images in the generation.

The results are collated in figure 3(b). The findings are inconclusive, partly because the two diversity measures often didn't agree. In earlier experiments, we found that, subjectively, the CLIP diversity estimation was more accurate than the Vendi score, and so we concentrate on that here. As expected, purely random generation and extending all the genomes in each round produces reliably diverse sets, with some exceptions. However, this gives no continuity or progression from generation 1 to 5, which can be useful in showing users how prompts and images evolve. Indeed, an initial motivation for introducing more sophisticated novelty searches was to have slightly less diversity in order to increase continuity. Of the non-random approaches, the CLIP clustering mechanism appears to have performed the best, with the CLIP novelty approach also performing well.

A reasonable use-case for the semantic search is for a user to choose a single image and then evolve it (keeping the seed) over five generations, using CLIP-guidance with respect to a target text. For each of the content texts above, and each of these targets: *fiery*, *abstract*, *minimal*, *yellow* and *psychedelic*, we simulated this use case 4 times. For comparison, we did likewise but using random choice rather than the semantic choice. At each generation, we recorded the highest CLIP similarity between an image and the target text for both semantic and random search. The average CLIP similarities over the sessions are recorded in figure 3(c). As expected, the semantic search always produced CLIP similarities (on average) higher than the random search. An example where the semantic search worked well is given in figure 2. Note that the random session is given down the left hand side. However, on inspection, we found that less than a quarter of the semantic searches produced images reflecting the target, hence there is much room for improvement.

Application to Artist Discovery

When Stable Evolusion is run in the Chrome browser, right-clicking any generated image allows that image to be used in a search for similar images, via Google Lens. In a small pilot study with two participants, we explored the potential for this to be used to discover artists that were previously unknown to the participant. The first participant was asked to imagine they were decorating a new apartment and wanted to find some human-painted physical artwork available for purchase online. They used Stable Evolusion for around 1 hour, starting with the content phrase 'downtown manhattan'. They produced 320 images over 40 generations, and searched online for physical artworks 17 times, each time finding something interesting from the traditional art world. They highlighted three artists whose work was particularly interesting, with details given in the appendix.

The second participant is an art historian, curator and cultural mediator with 15 years experience. They were asked to imagine a theme for a new exhibition and to use Stable Evolusion to find potential artists. The theme chosen was 'underwater world' and the session lasted 20 minutes, with the participant finding numerous potential artists, with 6 highlighted in the appendix. The participant pointed out that this approach helps break a chicken-and-egg problem in finding artists: it's hard to know in advance what to ask for a preliminary text search, but without such a search, images can't be found for image search to discover artists. With Stable Evolusion, they said, many different styles are offered to visualise the content text, some of which were new to them. Both participants expressed satisfaction in the ease of use of the system and the ability to discover artists. Both participants also noted that the user interface was cumbersome.

Conclusions and Future Work

Prompting image generators is currently a sought-after skill, and numerous approaches have been developed to automate prompt engineering. In (Martins et al. 2023), the authors implemented a similar evolutionary approach to prompt discovery as ours, but focused on quality of the results and match to user preferences, rather than artistic visualisations. Also, reverse engineering images to suggest prompts that will produce similar images is available via the CLIP Interrogator (huggingface.co/spaces/pharma/CLIP-Interrogator), and numerous other tools such as the Prompt Builder (promptomania.com/stable-diffusion-prompt-builder) are available to help in writing prompts.

We presented here the Stable Evolusion system which helps users find artistic visualisations of chosen content material via an evolutionary search which constructs prompts for text-to-image generation. We experimented to understand better how prompt evolution affects the images generated, and to evaluate the automated novelty and semantic searches. We demonstrated that the approach can help people to find human artists that they perhaps would not be able to through standard search methods. We plan to improve the user interface (moving to a HuggingFace Space), the search strategies, possibly using crossover techniques, and to increase functionality in discovering human artists.

Demonstration

The Stable Evolution colab notebook is available here:

`https://colab.research.google.com/drive/17sqwISmLbcpw3DEzMSzw1mbd8IlXBK4Z`

Acknowledgements

We would like to thank Stability AI for providing £500 of free credits for the DreamStudio API, enabling much preliminary experimentation. We also thank the anonymous reviewers for the helpful and insightful comments. Amy Smith and Sebastian Berns are funded by the IGGI centre for doctoral training [UKRI grant EP/S022325/1].

References

Bisong, E. 2019. Google colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress.

Colton, S.; Smith, A.; Berns, S.; and Murdock, R. 2021. Generative search engines: First experiments. In *Proceedings of the International Conference on Computational Creativity*.

Conditt, J. 2017. Google lens is a powerful, AI-driven visual search app. *Engadget, May 17*.

Friedman, D., and Dieng, A. B. 2022. The vendi score: A diversity evaluation metric for machine learning. arxiv 2210.02410.

Martins, T.; Cunha, J. M.; Correia, J.; and Machado, P. 2023. Towards the evolution of prompts with MetaPrompter. In Johnson, C.; Rodríguez-Fernández, N.; and Rebelo, S. M., eds., *Proceedings of EvoMusArt: Artificial Intelligence in Music, Sound, Art and Design*.

McCormack, J.; Cruz Gambardella, C.; Rajcic, N.; Krol, S.; Llano, T.; and Yang, M. 2023. Is writing prompts really making art? In Johnson, C.; Rodríguez-Fernández, N.; and Rebelo, S. M., eds., *Proceedings of EvoMusArt: Artificial Intelligence in Music, Sound, Art and Design*.

Patil, S.; Cuenca, P.; Lambert, N.; and von Platen, P. 2022. Stable diffusion with diffusers. *Hugging Face Blog*. huggingface.co/blog/stable_diffusion.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. arxiv 2103.00020.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-resolution image synthesis with latent diffusion models. arxiv 2112.10752.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Williams, Z. 2023. 'AI isn't a threat' – Boris Eldagsen, whose fake photo duped the Sony judges, hits back. *Guardian, 18th April 2023*.

Appendix

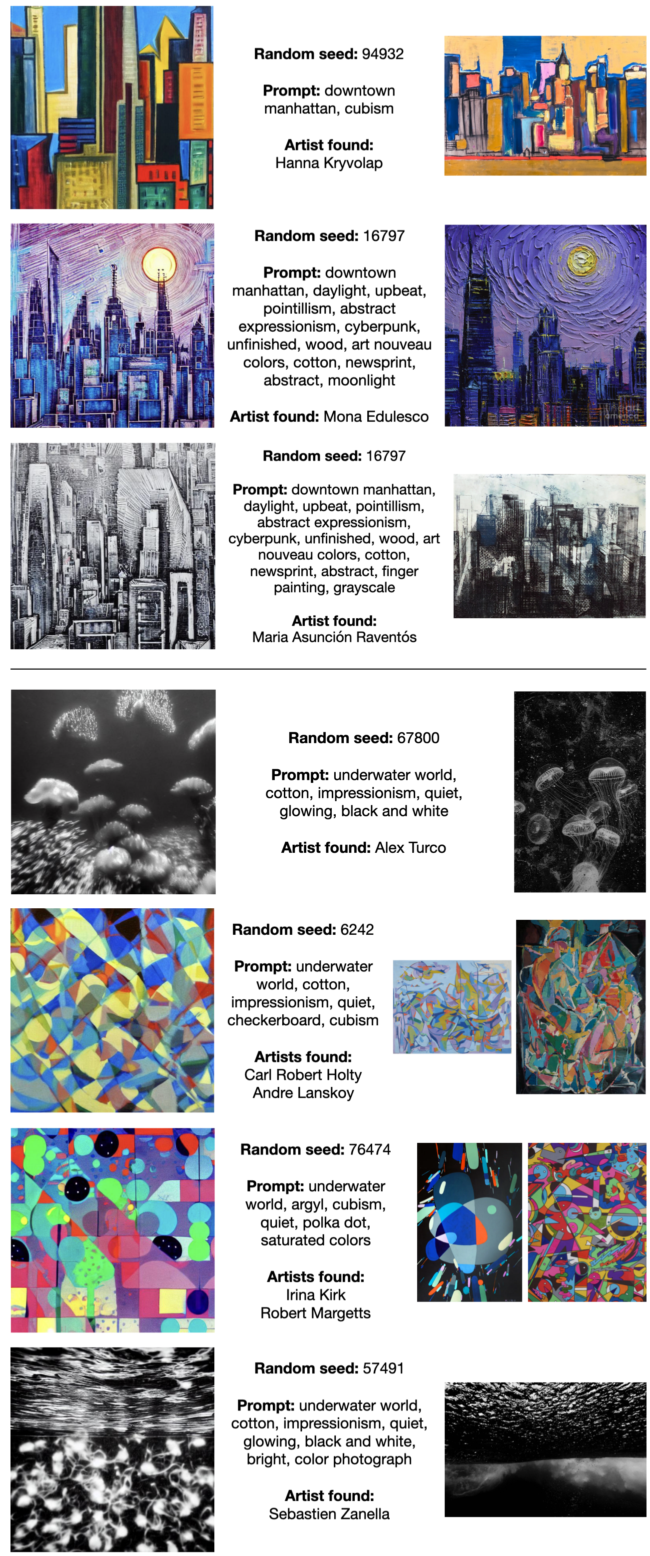


Figure 4. Example artist discoveries by study participants 1 (top) and 2 (bottom). Left: generated image; Centre: seed, prompt and discovered artist(s); Right: retrieved image(s).

Integrating AI/ML Techniques in Parametric Modeling for Creative Design

Jaechang Ko, Sora Key
Department of Architecture
Texas A&M University
College Station, TX 77845, USA
jaechangko@tamu.edu, sora@tamu.edu

Abstract

Parametric modeling is a fast way to create design variations but has limitations such as subjective parameter selections and constrained design variables. This study enhances parametric modeling by integrating AI/ML techniques, fostering creativity and innovation. The authors use deep learning algorithms to analyze 2D chair images, extract latent features, and employ t-SNE for visualization and parametrization of the features in 3D models. We also use 3D Generative Adversarial Networks (3DGANs) and conversational AI (ChatGPT) as design tools for novel chair designs, expanding design possibilities. This study demonstrates the potential for innovative design solutions, transforming the design process and suggesting new research directions.

Introduction

Professionals in creative domains, such as art, design, and architecture, have used parametric modeling to synthesize a multitude of design solutions by adjusting parameters during the modeling process, especially in trade-off relationships (Cross 2021). This method can rapidly generate a vast set of design variations, allowing designers to explore different possibilities, to define design problems more accurately, and to explore the opportunities and limitations of potential solutions (Schumacher 2015).

Despite its advantages, parametric modeling in the design area has several limitations. One limitation is the 'subjectivity' involved in parameter selection, as it depends on the designer's experience, knowledge, aesthetics, and personal preferences (Krish 2011). This subjectivity can often act as a double-edged sword: it provides a human's creativity and design uniqueness, but on the other hand, it could inadvertently limit the range of design possibilities explored (Menges and Ahlquist 2011). This is because individual biases, conscious or unconscious, could restrict the designer's perspective and the parameters chosen, potentially overlooking novel or unconventional design solutions that lie outside the designer's habitual thinking patterns. Another limitation is that the design variations generated by parametric models are inherently constrained by parameter interpolation. Iterative design exploration in the parametric modeling process focuses on individual parameters, rather than examining their inter-relationships (Yamamoto and Nakakoji 2005). As a result, entirely new designs cannot be created once the parameters are set, leading to a limited range of design possibilities and restricting the exploration of novel design solutions.

To address these limitations and enhance the design process, this study explores the integration of AI/ML techniques with parametric design methods to foster creativity and innovation in parametric modeling. The study aims to generate unique, dynamic, and innovative designs by using deep learning (DL) algorithms to analyze 2D chair images and extract latent feature space. Following this, we employ a dimensionality reduction algorithm (t-SNE) to visualize data distribution across the feature space, which serves as visual feedback for constructing parameters in the 3D model. This alternative approach allows for a more comprehensive exploration of the design space and facilitates the generation of novel design solutions, complementing human creativity and circumventing potential cognitive limitations in the design process. .

Furthermore, to expand the range of design possibilities beyond the constraints of parametric models, this study employs Generative AI models, 3D Generative Adversarial Networks (3DGANs), to create new 3D chair designs from the dataset generated by the parametric model. This approach allows for the generation of completely new design forms that are not limited by the initial parameter interpolation, thereby enabling greater design innovation.

This paper also explores conversational AI (ChatGPT) and its potential in design processes. This study demonstrates the unexpected generation of shapes and design solutions that are not reliant on predefined parameters. This approach offers a new method for designers to interact with design tools and discover unconventional design possibilities.

This study contributes to the existing body of research by bridging the gap between AI/ML techniques and parametric design methodologies. We demonstrate the potential for creating novel design processes and innovative design solutions that combine the strengths of both AI/ML and design methodologies through the application of deep learning algorithms, generative AI models, and conversational AI ChatGPT.

The paper is organized into the following sections: Section 2 describes the AI/ML augmented parametic modeling

ISBN: 978-989-54160-5-9

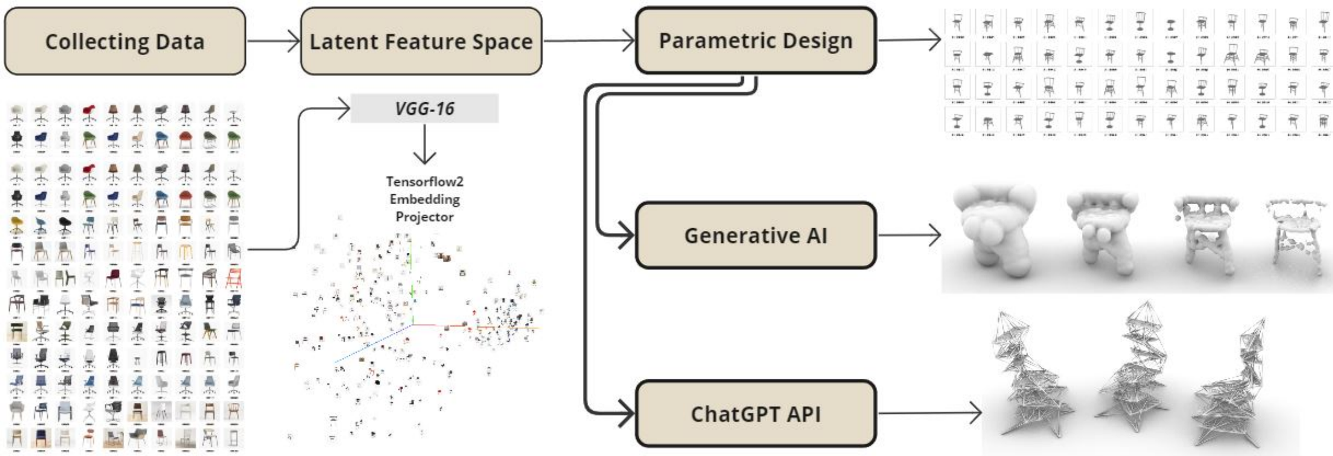


Figure 1: The overall process of parametric modeling and interaction with AI/ML technologies

process and illustrates how a feature space can contribute to design exploration and generation. Section 3 and 4 discuss the 3DGANs design and ChatGPT API implementation processes. Then, the advantages and limitations will be discussed with regard to integrating AI/ML techniques with parametric design methods, addressing the potential future research. (Figure 1)

Augmented AI/ML in Parametric Modeling

To conduct a manageable modeling process and meaningful geometric exploration, we chose chairs as the subject due to their significance in design history and their wide-ranging variety in form, function, and style (Cranz 1998). We randomly selected 300 chairs from the book '1000 chairs' by Charlotte & Peter Fiell (Fiell and Fiell 1997), ensuring a balanced representation of the diverse range of styles, including mid-century, Scandinavian, Brazilian, and others, in order to minimize bias in our dataset. The images were preprocessed to a resolution of 512x512 pixels, resulting in a feature vector of size 262,144.

Design Feature Space Generation

For the extraction of feature vectors, we utilized the VGG-16 model (Simonyan and Zisserman 2014) since the VGG-16 model has been widely adopted and proven effective in various applications, making it a reliable choice for our study (He et al. 2016). Additionally, this paper focuses on exploring the integration of AI/ML techniques with parametric design methods rather than achieving higher accuracy in feature extraction.

After obtaining the feature vectors, we used t-SNE, a dimensionality reduction algorithm (Van der Maaten and Hinton 2008), in TensorFlow's Embedding Projector to visualize the distribution of these features in a lower-dimensional space (Smilkov et al. 2016). This visualization allowed us to identify and analyze distinct clusters of chair designs. (Figure 2)

Parametric modeling process

There are no previous approaches to extracting the latent design space features and incorporating them into the para-

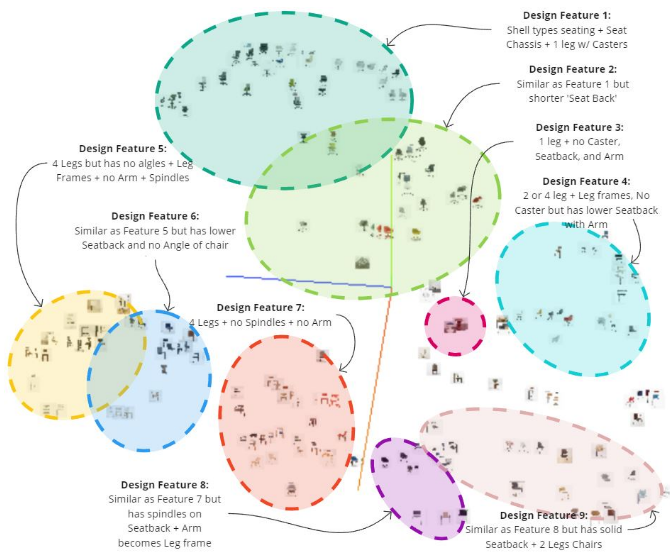


Figure 2: Latent space analysis to extract design features

metric design process. We extracted 9 distinct design features through the Embedding Projector and captured key aspects of chair design: Seat_shell, Seat_seat, Seat_back, Arm, Seat_base, Seat_chassis, Leg_frame, Leg, and Caster. These features were identified through careful analysis by practice experienced designers and faculties (Maxwell 2012).

Manual feature extraction, in particular, allows for incorporating domain knowledge into the design process. Therefore, designers can leverage their expertise to identify and select the most relevant features.

This methodology reflects the perspective of a designer and allows for the incorporation of human intuition and creativity in the AI/ML-augmented design process

To construct a parametric model, we organized design features using the following categories. The most prominent classification question started with "Does the chair have a shell seat or not?". For each category, different parametric rules were established based on factors such as the 'seat shape', 'number of legs', 'leg angles', and 'presence or absence of arm rest'. All generated 1000 chairs are shown in Figure 3.

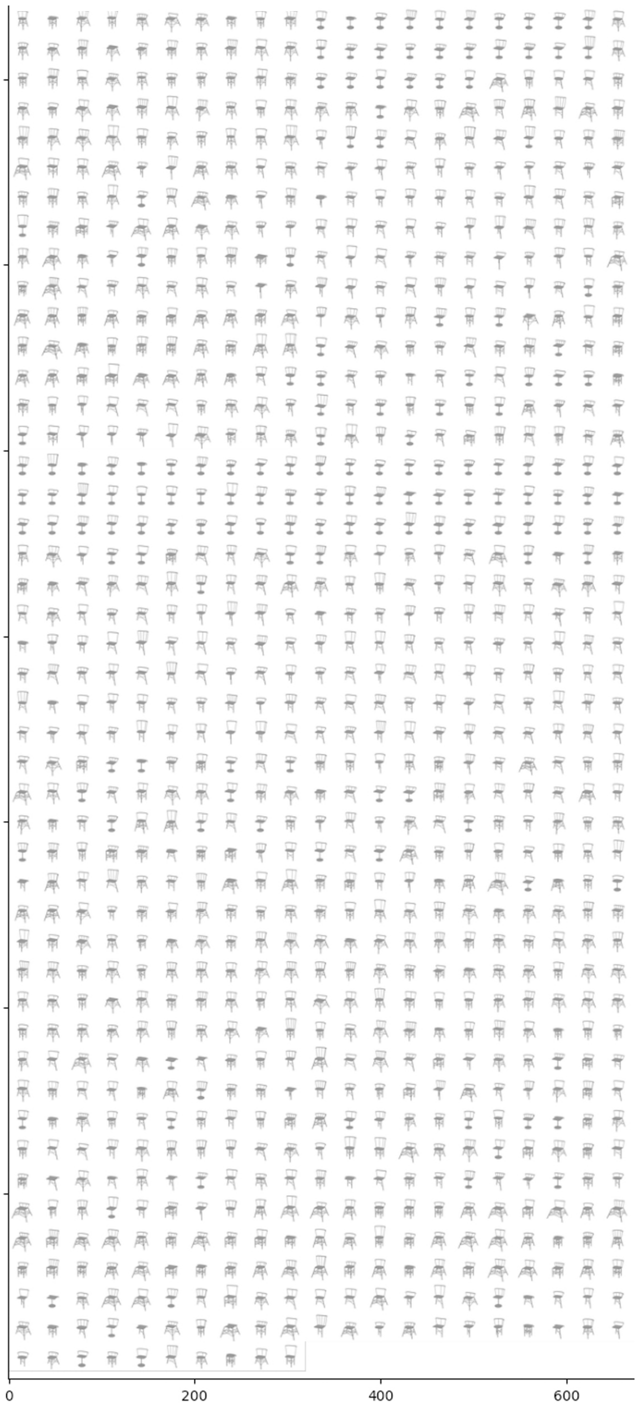


Figure 3: 1000 new chair designs resulting from parametric modeling

- Seat shape: We adjusted the 'fillet' parameters at all four corners, enabling the circular seat form generation. Moreover, when the arm's portions exceed the seat's radian value at a certain threshold, the parametric model decides that the chair has an integrated 'Arm' and 'Seat_back' together.

- Number of Legs: We accounted for a range of 1 to 4 legs. Implementing a Lloyd Algorithm, which repeatedly updates the centroid positions and assigns data points to their nearest centroids until convergence (Du, Emelianenko, and Ju 2006), we enabled the application of a bridge to the center point of each split plane. In cases where the seat shape is circular, we observed that the leg's center point falls outside the chair seat. To overcome the challenge of creating parametric models that cover all these scenarios, we derived a relational expression using a regression model, applying the position coordinate value to predict all different case scenarios.

- Leg Angle: Through feature space analysis, we verified that additional frames are braced when the leg angle surpasses a specific angle.

By assigning random values to each parameter, a thousand distinct chairs were generated, addressing the potential for innovative and unexpected outcomes when combining human creativity with computational power.

The AI/ML augmented parametric modeling process expanded the possibilities of the existing parametric design process by leveraging computational capabilities to identify, synthesize, analyze, and classify patterns and characteristics of the data, tasks which are extremely challenging for humans.

Generative AI/ML in Parametric Modeling

While parametric modeling is an iterative process that focuses on the variation of individual parameters and struggles with comparison and selection, 3DGANs can learn complex relationships among the parameters and generate new designs based on the distribution of data in the latent space (Zhang et al. 2019). In contrast to utilizing 3D geometry libraries such as ShapeNet (Chang et al. 2015), we trained the 3DGAN on a dataset of 3D models generated through the parametric modeling process. This approach allows the 3DGAN to learn the underlying structure and dependencies between the parameters, resulting in the generation of diverse and innovative design solutions that go beyond the predefined parametric space.

Data Preprocessing and Training for 3DGAN

In this experiment, the main goal was to extract features from 3D chair models rather than high-resolution data. The models were voxelized into 64-sized grids, resulting in a 64x64x64 representation. The entire dataset had a shape of [1,000, 32, 32, 32]. The 3DGAN was trained using a batch size of 64 and for 500 epochs. Once the generator successfully trained on the encoded datasets, the PyTorch tensor was converted to a NumPy array. To produce a clearer distinction between solid and empty spaces in the generated 3D chairs, a threshold of 0.5 was applied, converting the continuous-valued output to a binary voxel representation (1 for solid and 0 for empty).

Visualization and Design Exploration

We imported each voxel's center as a point coordinate information into Rhino Grasshopper. By manipulating the sphere's radius parameter, we were able to explore various chair forms derived from the generated points. (Figure 4)

By employing 3DGANs in the design process, we successfully expanded design possibilities beyond the inherent constraints of parameter interpolation. A latent space can be generated that is not confined to the parameter relationships found in traditional parametric models. It also allows for a more comprehensive exploration of design possibilities and facilitates the generation of novel design solutions that would otherwise be unattainable within the boundaries of parametric models alone.

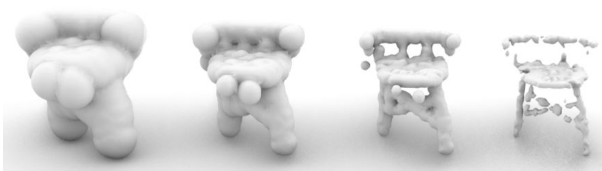


Figure 4: Re-parameterized 3D chair Generation

Conversational AI/ML in Parametric Modeling

We also explored the integration of ChatGPT, using Rhino Grasshopper (GhPython) code suggested by ChatGPT. This implementation aimed to facilitate an interactive design process where designers could engage in a dialogue with the AI model, allowing them to discover unconventional design processes, thereby creative solutions not confined to predefined parameters.

Conversation in the ChatGPT Website Platform

The first method uses the online ChatGPT platform as a conversational design assistant. Designers interact with ChatGPT by asking questions, discussing design ideas, and receiving generated GhPython code to create 3D shapes in Rhino Grasshopper. By doing so, the designers could continuously refine their ideas and solicit suggestions from the AI model. With those objectives in mind, the authors formulated the initial prompt as follows:

"Write a GHpython code that works in Grasshopper. Create a box-shaped geometry to define the outer boundary of the chair, then slice the box horizontally to create multiple layers. On each sliced surface, place a random number of points, then connect these points vertically & horizontally using polylines, with a maximum of 10 lines connected to each point."

By interacting with ChatGPT multiple times, the designer can continually refine the design idea and receive AI-generated suggestions to achieve the desired outcome. (Figure 5)

Integrating ChatGPT API into Grasshopper

We decided to create a bench using the ChatGPT API. Although chairs and benches share similar design attributes, they differ significantly in form and function, requiring a

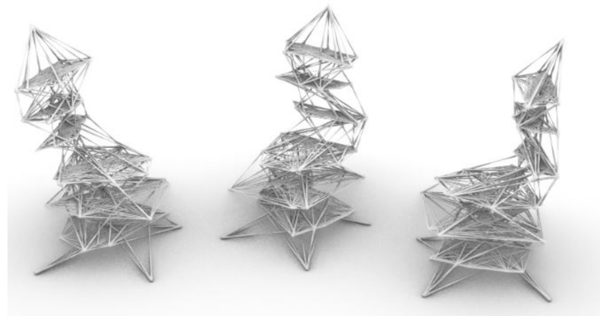


Figure 5: Chair design by Ghpython code from ChatGPT

creative approach to formal transformation. The conventional parametric design methodology, which primarily involves the combination of various parameters and variables, might constrain the scope of ChatGPT API, limiting its potential to transcend traditional design methods and templates. Thus, to fully utilize the capabilities of the ChatGPT API and simplify the design process, we developed a new modeling procedure that generates multiple cross-sectional curves using polynomial equations.

Firstly, we provided the following instructions to the ChatGPT API:

"Create two polynomials with three variables: x,y, and z. Seperate the equations with the symbol '&'. You can use sin or cos function"

Then, ChatGPT API generated the following such polynomials:

"$x^3 + 2x^2y + 3xyz \& cos(x)y^2z + y^3z^2 - 3x^2z^2$"

The authors subsequently utilized these polynomials to define points within the 3D space. These points were then connected to create smooth, curved lines. The authors then utilized these polynomials to define points within a three-dimensional space. These points were interconnected to form smooth, curved lines. The ChatGPT API repeatedly generated multiple curves every 5 seconds based on components and commands given to the grasshopper (See figure XX). The authors subsequently applied 'Loft' to create 3D geometry, which refers to creating a 3D surface or solid by interpolating between multiple 2D cross-sectional curves in Rhino. (Figure 6).

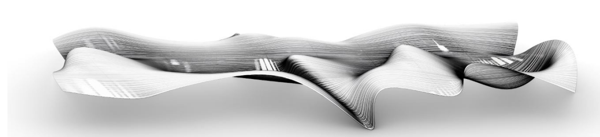


Figure 6: Bench design created by ChatGPT API

The integration created a unique 3D model that displays several features requiring further exploration and analysis. Especially the model presented a sculptural form that deviates significantly from the conventional understanding of a bench. Its design reminds the Verner Panton chair (Kim 2005), challenging conventional delineations of design elements such as the backrest, bench legs, and seat.

Conclusions

This study has successfully demonstrated the potential of integrating AI/ML techniques with parametric design methodologies to generate innovative and dynamic design solutions. Furthermore, this implementation process shows the possibilities to overcome the limitations of conventional parametric design approaches.

The integration of ChatGPT into the design pipeline remains a significant challenge. While ChatGPT has proven effective in devising new design methodologies, its integration with the design process could be further optimized. Currently, it functions as an independent component, but deeper integration could allow for more real-time, collaborative interaction between the designer and the AI.

We are currently developing the evaluation criteria for AI-generated designs for future work. We need to balance domain-specific aesthetics with broader, potentially domain-independent criteria for creativity. This will require careful thought and potentially novel methods for evaluating and determining the creativity and value of AI-generated artifacts. Moreover, user studies involving designers could provide valuable insights into the practicality and usability of the proposed techniques.

In conclusion, while this study has charted a new direction for AI in design, it represents only the beginning of an exciting journey. The challenges and limitations underscore the scope for future work in this domain. We hope this endeavor stimulates further research and development in integrating AI and design, transforming how we conceptualize, create, and evaluate design artifacts.

References

Chang, A. X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Cranz, G. 1998. *Chair: Rethinking Body Culture And Design*. WW Norton & Company.

Cross, N. 2021. *Engineering design methods: strategies for product design*. John Wiley & Sons.

Du, Q.; Emelianenko, M.; and Ju, L. 2006. Convergence of the lloyd algorithm for computing centroidal voronoi tessellations. *SIAM journal on numerical analysis* 44(1):102–119.

Fiell, C., and Fiell, P. 1997. *1000 Chairs*. Taschen.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kim, J. 2005. A study on the expressive characteristics in verner panton's chair design-focused on chairs designed from 1955 to 1970. *Korean Institute of Interior Design Journal* 14(2):178–187.

Krish, S. 2011. A practical generative design method. *Computer-Aided Design* 43(1):88–100.

Maxwell, J. A. 2012. *Qualitative research design: An interactive approach*. Sage publications.

Menges, A., and Ahlquist, S. 2011. *Computational design thinking: computation design thinking*. John Wiley & Sons.

Schumacher, P. 2015. Design parameters to parametric design. In *The Routledge Companion for Architecture Design and Practice*. Routledge. 35–52.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smilkov, D.; Thorat, N.; Nicholson, C.; Reif, E.; Viégas, F. B.; and Wattenberg, M. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*.

Van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(11).

Yamamoto, Y., and Nakakoji, K. 2005. Interaction design of tools for fostering creativity in the early stages of information design. *International Journal of Human-Computer Studies* 63(4-5):513–535.

Zhang, T.-F.; Tilke, P.; Dupont, E.; Zhu, L.-C.; Liang, L.; and Bailey, W. 2019. Generating geologically realistic 3d reservoir facies models using deep learning of sedimentary architecture with generative adversarial networks. *Petroleum Science* 16:541–549.

Co-creative Music Synthesizer Patch Exploration

Sean Luke and **Victoria Hoyle**
Department of Computer Science
George Mason University
Washington, DC, USA
sean@cs.gmu.edu
victoria.hoyle@protonmail.com

Abstract

We present an interactive evolutionary approach to exploring the space of synthesizer patches which combines an evolutionary optimizer with a variational autoencoder neural network. The objective is to work with musicians to explore the complex space of patches rather than program patches themselves, an often tedious and difficult task. The technique uses an algorithm to wander through the parameter space, while engaging the musician in assessing the quality of discoveries and providing real-time feedback to the algorithm. We describe the method and argue for it as a co-creative system.

Introduction

A *patch* is a program for a music synthesizer which directs it to produce a given kind of sound when played by a musician. The term dates from early synthesizers, which were programmed by connecting various modules with *patch cables* to control the flow of audio and modulation signals. Nowadays a patch is typically a fixed-length array of parameter values which together specify the nature of the sound generation elements used, their settings, and their connections.

Programming synthesizer patches can be daunting. While earlier synthesizers had relatively few parameters, modern synthesizers can have many hundreds of them. Indeed, some *additive synthesizers* have several thousand parameters, presenting a difficult high-dimensional design space. Still other synthesizers, such as *romplers*, have parameters consisting of many hundreds of unordered options. Critically, the relationships between parameters may be nontrivial. Some kinds of synthesizers, such as *subtractive synthesizers*, have parameters which are relatively independent of one another, and so their effect on the overall sound can be predicted and tuned independently. But other synthesizers, such as *frequency modulation (FM) synthesizers,* have parameters with strong and nonlinear relationships, and the impact of changing one parameter will strongly depend on the settings of others.[1]

[1] Some FM synthesizers were so difficult to program that musicians resigned themselves to playing only the factory patches which came on the units: these patches have since become famous. For example the Yamaha DX7's *E. Piano 1* and *Bass 1* factory patches were used on numerous pop songs, as was the Yamaha TX81Z's *LatelyBass* patch.

Finally many synthesizers, particularly those from the 1980s and 1990s, have very poor interfaces, making programming them from their front panels tedious.

But synthesizers do not have to be programmed only from their front panels: they can also be programmed via MIDI, a standardized serial port interface with a packet protocol. This makes it possible to design software tools called *patch editors* which allow the musician to program the synthesizer remotely using a better quality interface on a computer screen. However even with an improved interface, the number and complexity of synthesizer patch parameters can still make programming them a very difficult challenge.

An alternative is for the musician to collaborate directly with the patch editor in exploring the patch space. As it turns out, no less than Brian Eno proposed exactly this idea in a 1995 letter to Stewart Brand. He wrote:

> But what if the synthesizer just "grew" programs? If you pressed a "randomize" button which then set any of the several thousand "black-box" parameters to various values, and gave you sixteen variations. You listen to each of those, and then press on one or two of them — your favourite choices. Immediately the machine generates 16 more variations based on the "parents" you've selected. You choose again. And so on . . . . The attraction of this idea is that one could navigate through very large design spaces without necessarily having any idea at all of how any of these things were being made. I want to get some synth manufacturer interested in this. They are not too bright, in my opinion, so this might take a long time . . . . (Eno 1996) [p. 190].

In Luke (2019) we developed a method for doing this via interactive evolutionary optimization in Edisyn, a popular patch editor of our own design. Using this method, the editor wanders through the space of patches, discovering, proposing and auditioning ones to the musician, who assesses them. These assessments guide the editor in its search for new and better patches. In this paper we present an extension to this method which employs a combination of evolutionary optimization with a variational autoencoder trained on a large number of patches developed by the synthesizer community. In short, the method wanders not through the space of *all* patches, but through a manifold or subspace of patches which resemble, to some degree, the community patches themselves. We then discuss how and whether this back-and-forth between the program and musician is co-creative.

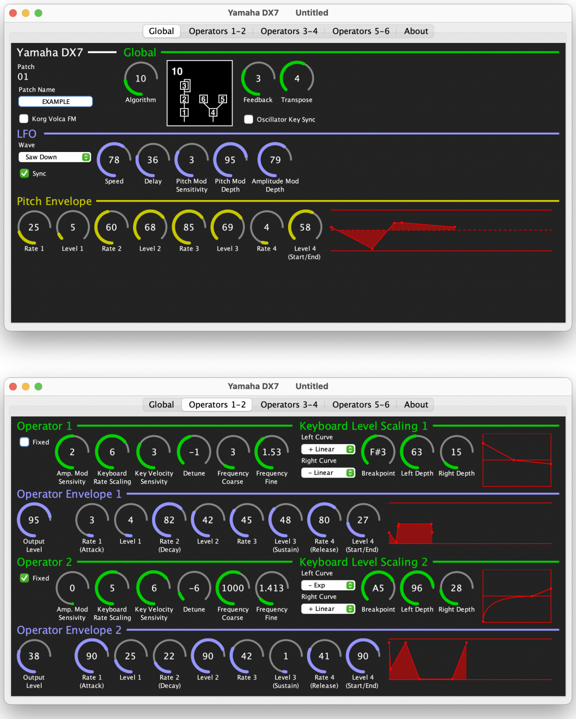


Figure 1: Edisyn's Yamaha DX7 patch editor, showing the "Global" and "Operators 1–2" panes.

Evolutionary Computation Evolutionary computation (or EC) is a family of stochastic optimization algorithms of which probably the most famous example is the Genetic Algorithm. An EC algorithm starts with a sample of randomly-generated candidate solutions (a *population* of *individuals*). It tests each individual according to some objective (or *fitness*) function. It then *breeds* a next-generation population by iteratively selecting and copying individuals from the previous population (the *parents*), recombining (mixing and matching) elements of the copies, and mutating the recombined copies with some degree of noise, producing their *children*. The selection procedure is biased to tend to select fitter individuals. Ideally over successive generations the current population improves in fitness. See Luke (2013) for more on EC.

Usually the fitness function is an automated procedure, but in our approach, the fitness of an individual is computed by auditioning the individual (the patch) in front of a human (a musician or sound designer), who offers an assessment. This approach is commonly known as *interactive evolution* and has been applied to a very wide range of fields ranging from art to robotics to industrial design (Takagi 2001).

Previous Work

The seminal paper in evolutionary patch optimization was Horner, Beauchamp, and Haken (1993), in which patches were proposed, played on the synthesizer, and then automatically compared for error against a target sound. Thus the fitness function was an automated procedure. This approach is known as *evolutionary resynthesis*.

Some later work has focused on interactive evolution, using a human to assess patch individuals. An early and well-known implementation of Brian Eno's original idea is *MutaSynth* (Dahlstedt 2001), a manual patch-recombination method which eventually found its way onto the commercial editor for the Nord Modular G2 synthesizer. The interactive evolution literature has considered different ways to deal with the difficulties inherent in auditioning patches, which take up time, for humans, who are easily bored. McDermott, O'Neill, and Griffith (2010) focused on interfaces designed to speed the assessment and selection of solutions. Seago (2013) simplified the search space by updating a parameterized model instead of a sample (essentially a form of *estimation of distribution algorithm*, see Luke (2019)). Suzuki et al. (2011) also simplified the search space by restricting candidate solutions to those drawn from an existing corpus of patches.

Rather than use neural networks in conjunction with evolutionary optimization as we have, some work has applied evolutionary computation in the *development* of neural-network-based synthesis methods (Ianigro and Bown 2016; Jónsson, Hoover, and Risi 2015).

Edisyn

Edisyn is a popular open source patch editor library of our design written in Java.[2] Edisyn has 76 patch editors supporting 139 synthesizers from 39 different families, plus editors for microtonal scales and for general MIDI parameter editing. These editors cover a wide range of synthesizer types: additive synthesizers, subtractive, rompler, drum, FM, and hybrid synthesizers; plus samplers, MIDI routers, and controllers. It attempts to present these using a unified and consistent interface. Figure 1 shows two of four panes from Edisyn's patch editor for the Yamaha DX7, a famous FM synthesizer.

Edisyn allows the musician to connect to a remote synthesizer over MIDI, then play notes on the synthesizer, change parameters in real-time, upload and download patches from the synthesizer's current working memory (the patch it is presently playing), read and write patches to the synthesizer's long-term patch storage, and load and save patches to disk on the musician's laptop. Edisyn also offers a *librarian*, essentially a spreadsheet of all patches on the synthesizer for bulk modification and organization.

Edisyn is distinguished among patch editors by its extensive set of automated patch exploration tools. This includes patch mutating, recombining two or many patches to form a child, "nudging" patches towards or away from other patches, and real-time morphing of patches as interpolations of up to four other patches. These features can be constrained in several ways, notably by restricting the parameters permitted to be mutated or recombined, and by specifying the degree of mutation or recombination involved. Prominent among the patch exploration tools is Edisyn's *Hill-Climber.*

The Hill-Climber

Edisyn's Hill-Climber is a patch space exploration tool using interactive evolution. It employs a variation of a so-called (μ, λ) Evolution Strategy algorithm with a highly customized

[2] Edisyn may be downloaded at https://github.com/eclab/edisyn

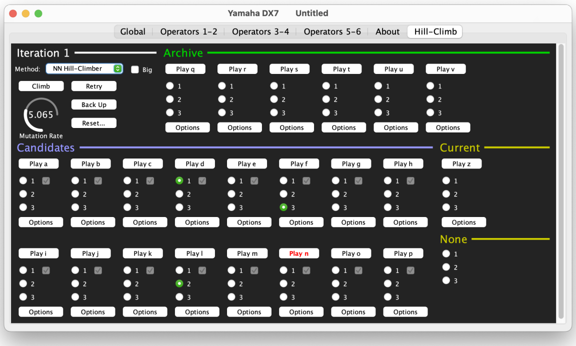


Figure 2: Edisyn's Hill-Climber in 16-Candidate mode (Variational Autoencoder turned on).

recombination and mutation method. For the Yamaha DX7 family of synthesizers this facility is further augmented with a deep-learned neural network (a variational autoencoder), as discussed later. The Hill-Climber, set up with the variational autoencoder, is shown in Figure 2.

The musician initializes the Hill-Climber by selecting a patch as a starting point. The Hill-Climber then seeds itself with 16 or 32 patches randomly selected from the vicinity of the initial patch. These patches are sent to the synthesizer to be auditioned to the musician one by one; the musician can request to re-audition a patch one at any time. The musician selects and ranks up to three patches as favorites. The Hill-Climber then *breeds* these patches to produce a new generation of 16 or 32 new patches in their vicinity. The new patches are auditioned to the musician and the process repeats.

At any time the musician can edit patches under consideration, save them, move them to other patch exploration tools, or back up to or build a new set of patches. The musician can designate a patch to be one of six "hall of champions" patches: any time later they may select and rank any "champion", as well as the current patch being edited, instead of an individual from the current generation. Finally, the musician can restrict the parameters that the algorithm is permitted to modify during optimization, and has control over the degree of mutation and noise applied at any time (and thus the balance between exploration and exploitation of the space).

The Hill-Climber employs an elaborate breeding mechanism which provides diversity and novelty while also offering patches which resemble ones preferred by the musician, as shown in Figure 3. The breeder relies on three mechanisms: mutation, recombination, and opposite-recombination. These algorithms are discussed in detail in Luke (2019), but we may summarize them here. Mutation adds noise to every parameter in a patch individual. If the parameter is metric, the noise is added by uniformly selecting from a range centered on the parameter value sized according to the musician's chosen mutation weight. If the parameter is categorical then its value is randomized with a certain probability again chosen according to the mutation weight.

Recombination takes two parent individuals and produces a child individual as follows. For each parameter in the first parent, with some probability the parameter will deviate from

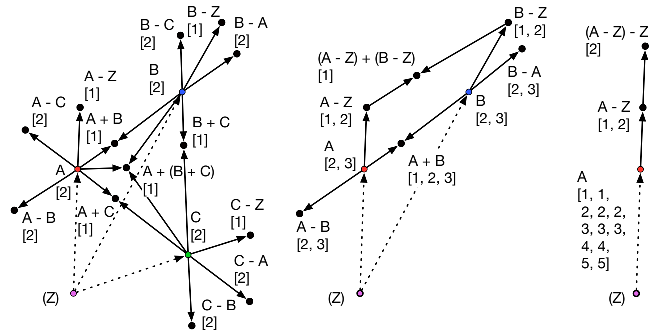


Figure 3: Breeding mechanism for the Hill-Climber when the musician has selected three parents (left), two (center), or one (right). A, B, and C represent these parents ranked best to worst by the musician, and Z is the previous generation's parent A. Children are produced near locations represented by nodes in the graph (other than Z). A child is produced by combining parents as shown ($+$ denotes standard recombination and $-$ represents opposite-recombination), then applying mutation. The notation $[a, b, \ldots, n]$ indicates the number of children produced at a location and their mutation counts. For example, $A+B$ [1, 2, 4] means that three children are produced by recombining A with B: one child is mutated once, one is mutated twice, and one is mutated four times.

the first parent's value. If the parameter is metric, the new value will be randomly selected from the range between the two parents. If the parameter is categorical, the new value will be, with 0.5 probability, set to the value of the second parent. Recombination is meant to "mix and match" features of two fit parents, ideally to produce yet fitter offspring.

Opposite-Recombination is a variant of Recombination meant to add diversity or act as an inertia procedure to push in the direction indicated by the musician's selections. For each parameter, it produces a value on the "other side" of the first parent from the second parent with some probability. If the parameter is metric, this is done by subtracting the first parent from the second. If the parameter is categorical, then the new value is set to the second parent unless they are the same, in which case it is set to some random different value.

Humans are a Problem The primary challenge in interactive evolution is the low number of individuals (patches) presented to the musician. It is common for an evolutionary optimization algorithm to require tens or hundreds of thousands of presentations before it has adequately optimized. This is not possible in interactive evolution, as the fitness function is a human, and humans are fickle, are easily distracted, and get bored quickly. It is not reasonable to expect a human to sit through more than a few hundred patch auditions before they give up. This difficulty is known as interactive evolution's *fitness bottleneck* problem (Biles 1994).

Because it has so few auditions available, the Hill-Climber must resort to tricks to maximize the value of each audition. The parameter space of patches is sparsely populated with "good" patches, and filled with garbage or silent ones, and the Hill-Climber must avoid these garbage patches. For example, the Hill-Climber's careful delineation of metric and categorical parameters, with custom mutation and recombination operators for each, avoids jumping into garbage space caused by treating all parameters as metric (as is commonly done).

This is also a reason for the unusual breeding mechanism: it does not deviate too far from the patches selected, but still enforces diversity and can provide an inertia mechanism: if the musician has moved from a previous patch Z to a new patch A, perhaps they would prefer a patch even further in that direction (see $A-Z$, $B-Z$, and $C-Z$ in Figure 3).

A version of the Hill-Climber called the *Constrictor*, employs a different garbage-avoidance strategy: starting from N well-vetted patches, it allows the user to iteratively remove patches, replacing them with recombined versions of the remainder. The idea is that by staying in the middle of a cloud of well-vetted patches, we are less likely to find garbage.

However perhaps the most aggressive approach to avoiding "bad" patches is the Hill-Climber's new, optional variational autoencoder, discussed next.

Variational Autoencoding

An *autoencoder* (Hinton and Salakhutdinov 2006) is a feed-forward neural network that takes an incoming vector and must output exactly the same vector. However, in the middle of the neural network there is a narrow neck through which data must pass. For example, the autoencoder might input vectors of length 100, but in order to output them must pass their information through a space of size 45. Obviously this cannot achieve an identity function in general: but it may be able to achieve the identity function on a finite training set of vectors. To do this it would learn a smooth, 45-dimensional latent space (a manifold) which passes through all of the training set vectors in the higher (100) dimensional space. It would then map incoming vectors to this latent space (or *encode* them) and then unmap them back again on the other side of the narrow neck (or *decode* them).

We employ a version called a *variational autoencoder* or VAE (Kingma and Welling 2014), which learns a distribution over the latent space instead of a direct mapping into it. This is achieved by having the middle layer of the autoencoder encode a collection of parameters that describe this distribution. As Gaussian distributions are commonly used and well understood, what is learned in the model for our method is a collection of means and standard deviations which describe separate Gaussian functions for each dimension. While learning, the network is penalized for deviating from standard normal distributions (in order to avoid collapsing into degenerate zero-deviation distributions) by using a weighted Kullback-Leibler divergence. The process during training is to encode the vector into the parameters for the distribution, sample from the Gaussian distribution as the latent vector, then decode this sampled latent vector. The variance inherent in this process helps map similar regions of the latent space to similar patches: if multiple vectors near each other in the latent space are sampled, and both are supposed to decode to the same final vector — as often happens during training as we will input the same vector many times — the network will ideally learn something about *the region of the latent space surrounding those vectors* in the decoder, and

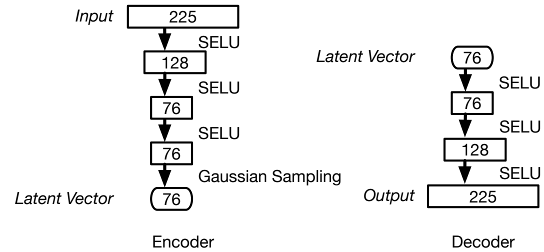


Figure 4: Encoder and Decoder Networks. Note that each block in the encoder and decoder describe their input size.

not just the vectors themselves. This architecture, known as β-VAE (Higgins et al. 2017), ideally restricts similar sounding patches, when manipulated in the latent space, to map to smooth, reasonable, and nearby patches in the final space.

After training, the VAE is broken into the *encoder*, which maps the full space into the latent space, and the *decoder*, which does the opposite. The encoder's Gaussian sampling layer is then replaced with the identity function. We can then, for example, input a random vector to the Decoder, and it would output a vector along the manifold defined by the original samples.

Improving Patch Optimization We train a VAE on a large corpus of human-designed patches. After training, we then separate it into the encoder and decoder. We primarily use the decoder as follows. The Hill-Climber is no longer maintaining a population of 16 or 32 patches: rather it is maintaining a population of *vectors in the latent space*. To assess a vector, it decodes it to produce the patch, then auditions it. To seed the initial generation, it simply uses the encoder to encode patch seeds into latent-space vectors for the population.

How does this help us? We first train the autoencoder on a large corpus of open, human-designed patches, and so the latent space only passes through the parameter space in the vicinity of these patches. Thus arbitrary vectors on the latent space will generally map to vectors in the regions populated by "good patches", avoiding garbage. Unfortunately, there are few synthesizers with an online corpus of enough patches to successfully train an autoencoder. The Yamaha DX7 is one: we have successfully trained an autoencoder using nearly 27K unique patches. The Yamaha TX81Z synthesizer is possible target candidate for the future, with approximately 8K patches available. The DX7 has 145 parameters, some categorical, and so when one-hot-encoded it comes to 225 parameters. The trained latent space was 76 parameters.

The specific architecture used can be seen in Figure 4. The *SELU* is an activation function which behaves identically to the well known *Exponential Linear Unit*, but has chosen scaling parameters which cause the weights to self-normalize over many training iterations (Klambauer et al. 2017).

We also use the autoencoder for simple patch mutation: rather than mutate a patch directly by some amount, we encode the patch into its latent vector, mutate the vector by that amount, then decode the result into a new patch. The goal, once again, is to mutate the patch but keep it near "reasonable" patches even with significant mutation weights.

Is this Computationally Co-Creative?

It seems clear that the Hill-Climber is at least a *creative support tool* in the sense of Shneiderman et al. (2005): it supports exploration, is forgiving of error, has a low threshold to entry, and is capable of exploring any part of the space. But this is a very low bar: it's the same for many very rudimentary tools. Instead, we argue that this tool is in fact co-creative.

Karimi et al. (2018) define computational co-creativity as "interaction between at least one AI agent and at least one human where they take action based on the response of their partner and their own conceptualization of creativity during the co-creative task." We think that the Hill-Climber easily achieves this: it is taking action (proposing new patches) based on the response of the human, and the human is taking action (criticism) based on the proposals.

The Hill-Climber is an asymmetric collaboration: it is proposing new patches, and while the musician *can* propose patches to consider, they are primarily the fitness function or critic. Thus we may view the Hill-Climber as a DIFI (or Domain Individual Field Interaction) system (Feldman, Csikszentmihalyi, and Gardner 1994). From a DIFI perspective, the Hill-Climbing algorithm is the Individual, and the musician is the Field (and, if you like, the Domain).

However, a creative system must typically optimize for two criteria at once: novelty and some notion of value (Boden 1992; Wiggins 2006). What is the Hill-Climber really optimizing against? After all, the fitness function is being entirely determined by a human being. It's true that the system is emphasizing both diversity (if not novelty) and quality when breeding, but ultimately it ought to be considered co-creative only if the *human,* in collaboration, is also aiming for novelty and value when assessing fitness. We imagine that this is the case in many situations: but humans are fickle. The system as a whole is co-creative, in some sense, only if the human is doing their part.

Conclusion and Future Work

We presented a system which combines interactive evolution and a variational autoencoder to help explore the space of synthesizer patches. We think this back-and-forth qualifies it as a co-creative system: or certainly something rather more than just a creative support tool.

The biggest challenge in interactive evolution still remains the fitness bottleneck. To progress even faster, we'd need to allow supervisory feedback: that is, allowing "I'd like the sound brighter" or "more like a cello" rather than just "I like this one better". This would make it easier to argue in favor of co-creativity as well: as the musician would be able to contribute more to the system than mere criticism.

References

Biles, J. A. 1994. GenJam: a genetic algorithm for generating jazz solos. In *International Computer Music Conference*, 131–137.

Boden, M. 1992. *The Creative Mind*. Abacus.

Dahlstedt, P. 2001. A MutaSynth in parameter space: interactive composition through evolution. *Organized Sound* 6(2):121–124.

Eno, B. 1996. *A Year With Swollen Appendices: Brian Eno's Diary*. Faber & Faber.

Feldman, D. H.; Csikszentmihalyi, M.; and Gardner, H. 1994. *Changing the World: A Framework for the Study of Creativity*. Praeger.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 303:504–507.

Horner, A.; Beauchamp, J.; and Haken, L. 1993. Musical tongues XVI: Genetic algorithms and their application to FM matching synthesis. *Computer Music Journal* 17(4):17–29.

Ianigro, S., and Bown, O. 2016. Plecto: a low-level interactive genetic algorithm for the evolution of audio. In *EvoMUSART*, 63–78.

Jónsson, B.; Hoover, A. K.; and Risi, S. 2015. Interactively evolving compositional sound synthesis networks. In *Genetic and Evolutionary Computation Conference*, 321–328.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. M. 2018. Evaluating creativity in computational co-creative systems. In *International Conference on Innovative Computing and Cloud Computing*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-normalizing neural networks. In *Neural Information Processing Systems*, 972–981.

Luke, S. 2013. *Essentials of Metaheuristics*. Lulu, 2nd edition. http://cs.gmu.edu/∼sean/book/metaheuristics/.

Luke, S. 2019. Stochastic synthesizer patch exploration in Edisyn. In *EvoMUSART*.

McDermott, J.; O'Neill, M.; and Griffith, N. J. L. 2010. Interactive EC control of synthesized timbre. *Evolutionary Computation* 18(2):277–303.

Seago, A. 2013. A new interaction strategy for musical timbre design. In *Music and Human-Computer Interaction*. Springer. 153–169.

Shneiderman, B.; Fischer, G.; Czerwinski, M.; Myers, B.; and Resnick, M. 2005. *NSF Workshop Report on Creativity Support Tools*. National Science Foundation.

Suzuki, R.; Yamaguchi, S.; Cody, M. L.; Taylor, C. E.; and Arita, T. 2011. iSoundScape: Adaptive walk on a fitness soundscape. In *EvoApplications*, 404–413. Springer.

Takagi, H. 2001. Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE* 89(9):1275–1296.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19:449–458.

Latent Lab: Large Language Models for Knowledge Exploration

Kevin Dunnell[2], Trudy Painter[1], Andrew Stoddard[1], Andy Lippman[2]
[1]Department of Electrical Engineering and Computer Science, MIT
[2]MIT Media Lab
{tpainter, dunnell, apstodd, lip}@mit.edu

Abstract

This paper investigates the potential of AI models, particularly large language models (LLMs), to support knowledge exploration and augment human creativity during ideation. We present "Latent Lab" an interactive tool for discovering connections among MIT Media Lab research projects, emphasizing "exploration" over search. The work offers insights into collaborative AI systems by addressing the challenges of organizing, searching, and synthesizing content. In a user study, the tool's success was evaluated based on its ability to introduce users to an unfamiliar knowledge base, ultimately setting the groundwork for the ongoing advancement of human-AI knowledge exploration systems.

Introduction

The untapped potential of collective knowledge holds significant implications for idea evolution and innovation across various entities (Curley and Salmelin 2013). Despite the digital revolution, information organization remains strikingly similar to traditional methods, limiting exploration across diverse sources and impeding the discovery of interconnected relationships. Current search approaches prioritize quick answers and display results in a list format. This hinders the discovery of interconnected relationships required for meaningful exploration and undermines the context of search terms by prioritizing keywords over semantics.

In contrast, synthesis tools like ChatGPT[1] offer a paradigm shift in user interface design through conversational interaction, though they have drawbacks such as the opaqueness of information sources and limited text-based interaction. This paper outlines the development of Latent Lab[2] and evaluates it in the context of the MIT Media Lab data set of 4,000+ research projects. This exploration tool transcends previous search and synthesis tools by incorporating browsing and active visual interaction. Leveraging data manipulation libraries, interactive visuals, and LLMs, Latent Lab overcomes the constraints of keyword-centric search, allowing users to engage in semantically meaningful exploration and synthesis of large data sets. The iterative design process of the tool itself highlights the importance of exploration in the creative process, offering a glimpse into the potential of AI-assisted idea generation.

We make the following contributions to the field of human-AI interactive knowledge exploration systems.

- We present the design and implementation of an interactive knowledge visualization tool, including a novel automated technique to label idea clusters using an LLM.
- We report the results from a user evaluation study, demonstrating the utility of a hybrid search/synthesis system to find meaningful insights and connections often overlooked by traditional search and synthesis tools.

[1] `https://chat.openai.com/`
[2] Try Latent Lab at `https://latentlab.ai/`

Related Work

Knowledge Organization

Vannevar Bush's memex laid the foundation for hypertext and associative indexing (Bush 1945). Richard Feynman's triangulation method emphasized understanding relationships between concepts (Feynman, Gottlieb, and Leighton 2006). These ideas influenced the development of Google Knowledge Graph (Carr 2007). Our approach to knowledge organization builds on these works to enable fluid exploration of linked information.

Information Visualization

Shneiderman's taxonomy established information visualization principles, with the "overview, zoom and filter, details-on-demand" mantra guiding the design of visual interfaces for interacting with large data sets. (Shneiderman 1996). Bostock et al. presented D3.js for interactive visualizations (Bostock, Ogievetsky, and Heer 2011). Heer and Shneiderman highlighted the importance of interaction in visual analysis (Heer and Shneiderman 2012). Our work integrates these principles to create an informative interface for users.

Information Retrieval

Spärck Jones introduced the term frequency-inverse document frequency (TF-IDF) weighting scheme for keyword-based search (Spärck Jones 1972). Mikolov et al. proposed the Word2Vec model for embedding-based search (Mikolov et al. 2013). Devlin et al. developed BERT, which further improved semantic search (Devlin et al. 2018). Latent Lab extends this work, using embedding-based search for relevant results in complex data landscapes

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

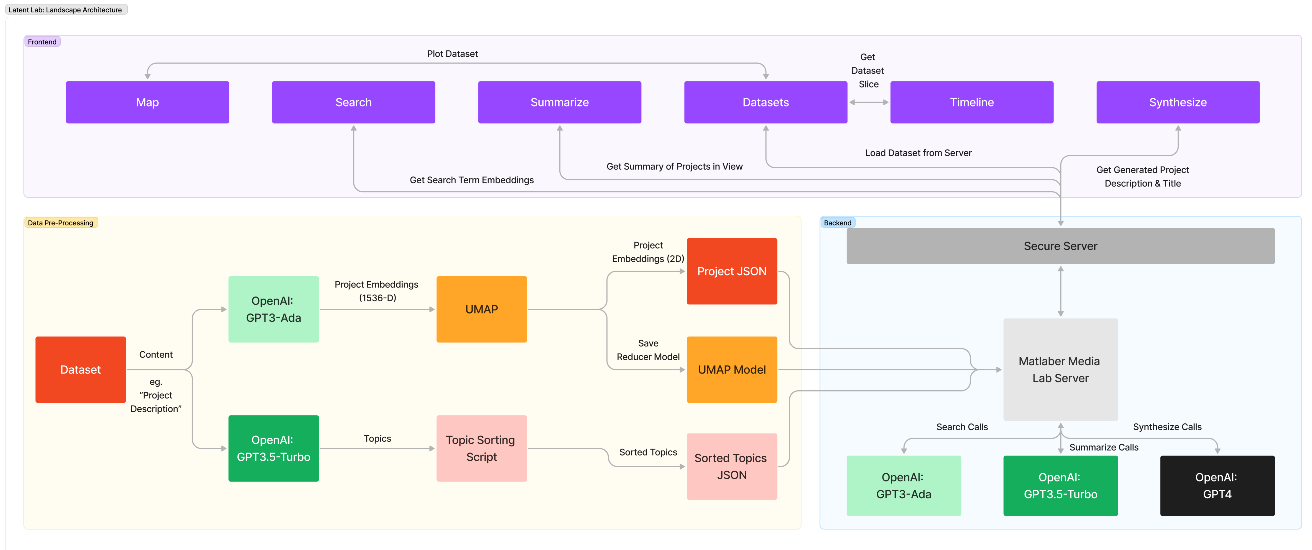


Figure 1: System Architecture of Latent Lab

Human-AI Collaboration

Minsky's Society of Mind proposed human intelligence as a result of interacting agents (Minsky 1988). Influential works that consider humans and intelligent systems as interacting agents include TRIZ, Polya's work on invention, and Weis and Jacobson's DELPHI framework (Weis and Jacobson 2021; Polya 1945; Altshuller 1999). Our work further examines human-AI collaboration, aiming to create a system that amplifies human capabilities and positions AI as a "copilot" rather than an "autopilot."

Methods

System Overview

Latent Lab is an AI-powered knowledge exploration system. High-dimensional unstructured data is condensed and visualized in an interactive 2D map. The interface allows users to explore labeled clusters of similar topics, search by semantic context, and synthesize new ideas.

System Architecture

Latent Lab's system architecture integrates state-of-the-art technologies. The back end is powered by Fast API[3] and Python, while the front is built with Vercel,[4], Next.js[5], React[6], and TypeScript. Initially, we aimed to execute all operations on the front end, but the lack of a fully JavaScript-ported version of UMAP (McInnes, Healy, and Melville 2018) necessitated the incorporation of a back-end server. This adjustment also enabled server-side rendering, significantly speeding up data loading. The system architecture diagram is presented in Figure 1.

[3]https://fastapi.tiangolo.com/
[4]https://vercel.com
[5]https://nextjs.org
[6]https://reactjs.org

Data Processing

The data processing pipeline is mostly automated and runs independently of the web app back end for each new data set. It generates three primary artifacts:

- A project JSON containing the unstructured data and embedding data for mapping every project on the front end
- A sorted research topics JSON containing all topics produced by the pipeline, ordered by topics with the most associated projects, used for the labels on the front end
- A pickled UMAP model to reduce project and topic embeddings to 2 dimensions on the back end

Topic Extraction

Latent Lab's automated topic extraction feature sets it apart from other embedding visualization tools, which don't provide insights into cluster meanings. The system uses GPT-3.5-Turbo to distill topics for each project, count occurrences of unique topic labels, and identify related projects. It then calculates label positions using the centroid of the UMAP-reduced coordinates for each associated topic.

Components

The Latent Lab interface has four main components, shown in Figure 2. It includes a Map Visualization, Generation Workbench, Search Bar, and Timeline Slider.

Map Visualization

The main visualization displays an organized map of project data, with dots representing research projects and clusters indicating semantic similarity. Dot colors correspond to different Media Lab research groups and can be customized to represent other discrete data set attributes.

Contour lines in the map indicate data density within clusters, a concept borrowed from topographic maps where they

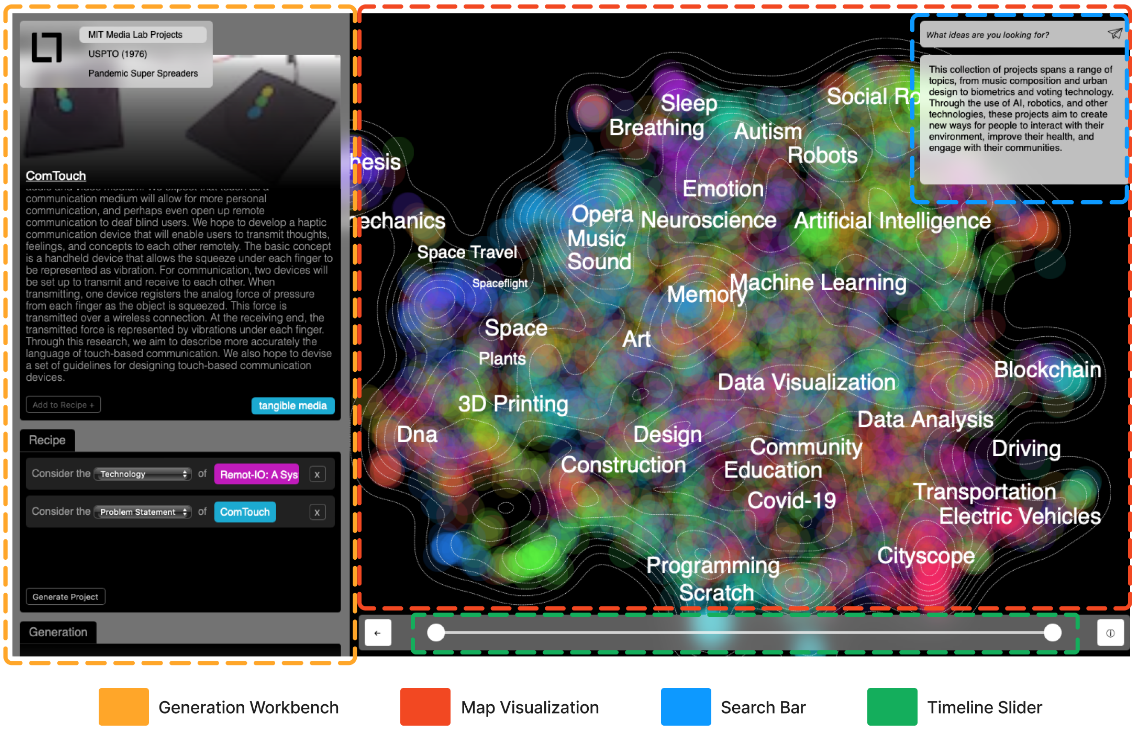


Figure 2: Latent Lab Interface, Annotated to Differentiate Between Components

represent elevation. Paired with the timeline, the changing contour lines reveal the evolution of research concentration.

Users can pan and zoom, uncovering varying levels of information. High-level labels and contour lines are shown at the highest zoom level, while sub-topic labels and project details appear when zooming in. An occlusion algorithm determines label visibility based on popularity and bounding box overlap.

Generation Workbench

Latent Lab's Generation Workbench allows users to create a "recipe" for collaboratively synthesizing new research project ideas. Users can choose whole projects or specific aspects, such as community, problem statement, or technology, to include. Once a recipe is prepared, selecting "generate" submits a preset prompt with selected project elements to GPT-4 via the OpenAI API, producing a synthesized project title and description. Users can view the exact prompt by clicking the "What was used to generate this?" information button. See Figure 3 for the user flow diagram.

Search & Summarization

Latent Lab employs embedding-based search for semantic meaning instead of simple keyword-matching, enabling more intuitive project exploration through contextual relationships. When a user searches, the query is sent to the back-end server, and the GPT-Ada API returns a 1,536-value embedding. This is passed to the UMAP reducer, yielding x and y coordinates, which are sent to the front end to zoom and highlight the relevant map region dynamically. Figure 4 demonstrates this process using "quadratic voting" as the search term. Below the search bar, Latent Lab displays summaries that give users a quick overview of projects in the currently viewed map region.

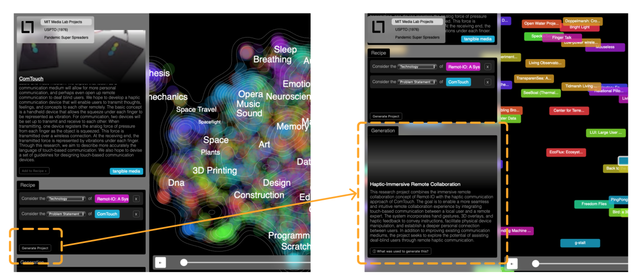


Figure 3: Generating a Research Project Idea

Timeline Slider

Latent Lab's Timeline Slider enables users to explore data set progression over a selected period using start and end date sliders. This functionality, particularly useful alongside the search bar, allows for efficient examination of current or ongoing projects in specific areas. Figure 5 illustrates timeline filtering for projects since 2018.

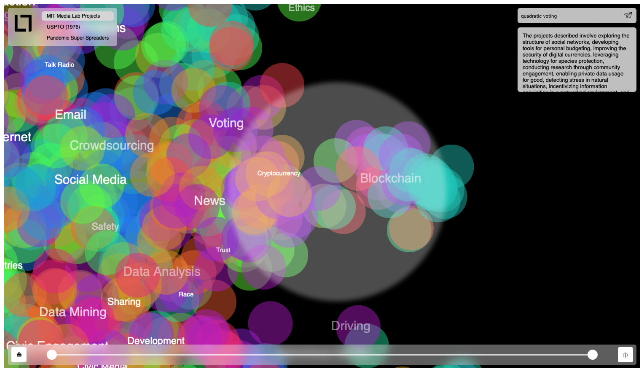


Figure 4: Search Highlighting in Map

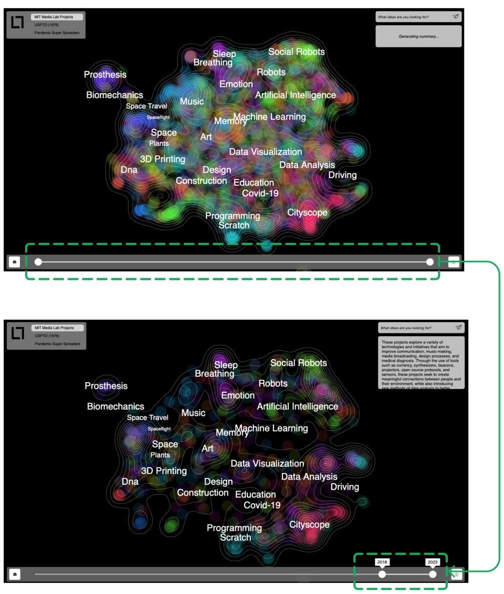


Figure 5: Timeline Evolution

Study Overview

We designed a study to evaluate users' experience exploring MIT Media Lab research using Latent Lab. The study compared Latent Lab to the current MIT Media Lab website, which uses traditional keyword-based search. Surveying 94 self-identified researchers via Prolific, participants interacted with both tools in a randomized order. After using each tool, they answered questions assessing clarity, effort (Hart and Staveland 1988), engagement, mental support, future use, trust (benevolence, capability, and reliability) (Mayer, Davis, and Schoorman 1995), and insight on a 1-5 Likert scale. The study aimed to measure Latent Lab's effectiveness in fostering human-AI collaboration, enhancing user experience, and promoting a deeper understanding of Media Lab projects with AI-powered tools.

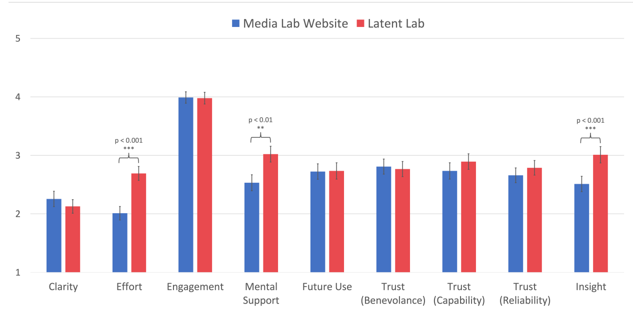


Figure 6: User Evaluation Results

Results

Analysis

The study results in Figure 6 indicate that Latent Lab shows promise as an AI-assisted exploration tool compared to the Media Lab website. Participants were equally engaged, trusted both systems and expressed equal likelihood to use them in the future.

Although Latent Lab required more effort, this is likely due to its novel functionality compared to traditional search interfaces. As users become more familiar with the design, we expect this effort to decrease, facilitating seamless human-AI collaboration. Latent Lab outperformed the Media Lab website in providing higher mental support and insight, suggesting that its semantic map effectively organizes knowledge and offers a deeper understanding of MIT Media Lab research than the current Media Lab website.

Overall, the study highlights Latent Lab's potential and underscores the need for minor improvements to deliver a consistent user experience and enhanced search results.

Future Directions

While our initial inspiration was drawn from our system's ability to generate research project ideas, early user feedback underscored the need to refine Latent Lab's knowledge organization for enhanced exploration, which took precedence over a thorough evaluation of the generated ideas. Looking ahead, our research will adopt a two-fold approach. Firstly, we aim to conduct a comprehensive evaluation of our tool's creativity as an ideation system, benchmarking the utility, novelty, and feasibility of the generated ideas. Secondly, we intend to enhance system performance for handling large user-uploaded datasets and improve data navigation and usability. This will necessitate a focused study on data visualization techniques to optimize Latent Lab's usability, with the ultimate goal of reducing user effort and maximizing the tool's potential for insight extraction.

Conclusion

Latent Lab serves as an innovative and powerful tool for exploring interconnected relationships within large data sets. By utilizing LLMs and visually engaging interfaces, it transcends conventional search limitations, providing a semantically meaningful and context-aware experience. Emphasizing the value of exploration and iterative design, Latent Lab realizes the long-sought goal of information technology experts for an intuitively accessible wealth of interconnected information. AI-assisted exploration has turned this vision into reality, setting the stage for future human-AI co-invention systems and fostering more intuitive and productive collaborations that are capable of generating novel and impactful creations.

Author Contributions

KD is lead author, TP is second author and contributed to writing and editing. KD, TP, and AL contributed to the conception of Latent Lab. KD, TP, and AP all contributed to the development of Latent Lab.

Acknowledgments

We thank Joshua Bello, Deborah Jang, Michael Peng, Jinhee Won, and Anushka Nair for their help with Latent Lab development. We also thank Ziv Epstein and Hope Schroeder for their guidance with user testing. Finally, we thank all survey participants for their time.

References

Altshuller, G. 1999. *The Innovation Algorithm: TRIZ, systematic innovation, and technical creativity*. Technical Innovation Center, Inc.

Bostock, M.; Ogievetsky, V.; and Heer, J. 2011. D3: Data-driven documents. *IEEE transactions on visualization and computer graphics* 17(12):2301–2309.

Bush, V. 1945. As we may think. *The Atlantic Monthly* 176(1):101–108.

Carr, N. 2007. Freebase: A case study.

Curley, M., and Salmelin, B. 2013. Open innovation 2.0: a new paradigm. *OISPG White Paper* 1–12.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feynman, R. P.; Gottlieb, M. A.; and Leighton, R. 2006. *Feynman's Tips on Physics: A Problem-Solving Supplement to the Feynman Lectures on Physics*. New York: Basic Books.

Hart, S. G., and Staveland, L. E. 1988. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology* 52:139–183.

Heer, J., and Shneiderman, B. 2012. Interactive dynamics for visual analysis. *Communications of the ACM* 55(4):45–54.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20(3):709–734.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minsky, M. 1988. *The Society of Mind*. Simon and Schuster.

Polya, G. 1945. *How to Solve It: A New Aspect of Mathematical Method*. Princeton University Press.

Shneiderman, B. 1996. The eyes have it: A task by data type taxonomy for information visualizations. 336–343.

Spärck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1):11–21.

Weis, D., and Jacobson, R. 2021. Delphi: A framework for human-ai collaborative problem solving and learning. *arXiv preprint arXiv:2106.03852*.

Composing Mood Board with User Feedback in Concept Space

Shin Sano[1] and Seiji Yamada[2]

[1,2]Department of Informatics, Graduate University for Advanced Studies (SOKENDAI), Tokyo, Japan
[1]Institute for Creative Integration, Oakland, CA, United States
[2]Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo, Japan
Email: [1]ssano@creative-integration.com, [2]seiji@nii.ac.jp

Abstract

We propose the Mood Board Composer (MBC), which supports concept designers in retrieving and composing images on a 2-D concept space to communicate design concepts visually. The MBC enables users to search images intuitively. Its algorithm adapts the query vector for the next search according to the user's rearrangement of images on the 3×3 grid. The next image search is performed by obtaining the most similar words from the adapted mean vector of the images on the grid thus obtained and using them as a new query. Our participants' experiment with 211 cases of mood board creation confirmed the effectiveness of adaptive iterations by the Creativity Support Index (CSI) score.

Introduction

Mood boards are visual artifacts often used as design development tools to communicate and share design ideas, such as emotions, feelings, or "moods" between stakeholders(Lucero 2012). They are often used in design practice and education, such as thinking externalization, meaning acquisition, and conceptual reasoning (Li and Zhao 2021). Mood boards are also used as qualitative design research tools facilitating creative thinking, presenting and communicating products (Cassidy 2011), communicating the designers' imagination and ideas they are pursuing (Edwards, Fadzli, and Setchi 2009). Bouchard et al. (2005) discuss the role of mood boards as intermediate representations (IR) in design at different levels of abstraction.

A mood board-composing task involves a variety of algorithmic problems to solve, such as image retrieval, search strategies, computer vision, semantic feature engineering, natural language processing, and query expansion and modification based on user feedback. Setchi et al. (2011) attacked the problem of the semantic gap of content-based image retrieval, and proposed a semantic-based approach that relies on textual information around the target image to avoid low-level and literal labels from given images. The method extracts the most relevant words in a document utilizing TF-IDF and a general-purpose ontology to expand the queries to find more of relevant images. Koch et al. (2019) created an interactive digital tool to support designers in creating a mood board, utilizing exploration-exploitation strategy optimized by a cooperative contextual bandit reinforcement learning algorithm. They further advanced the digital mood board tool(Koch et al. 2020), utilizing Google's Vision API to assign semantic labels to each image. The above two studies incorporates user feedback while engaged in the mood board creation task. Yet, no prior research has used the positions of images on the mood board to adjust queries for new images, and assumed a semantic space model on a quadrant system on which designers can position their ideation relative to the Design Concept Phrases (DCP) (Sano and Yamada 2022) of a target design concept.

Method

The MBC is an AI-assisted interactive web application designed to be used by concept designers who wish to explore and communicate their design concepts visually. It also intends to build on the idea of the Character Space (CS) and the Design Concept Phrase (DCP) (Sano and Yamada 2022), on which users explore design concepts in a lexicosemantic space. Mood boards composed by MBC are constructed in a grid of $n \times m$ tiles. Although no prior research specifically found the optimum number of images on a mood board, a few recent studies have indicated that participants in their studies typically handle 5 to 12 images per mood board(Koch et al. 2020; 2019; Zabotto et al. 2019). Aliakseyeu et al.(2006) experimented with different sizes of digital image piles to compare human performances on navigation, repositioning, and reorganizing tasks and found significant differences in task performance between two different pile sizes (15 and 45). In our development and experiment, we chose nine images with a 3×3 grid to facilitate users in quick glancing and iterations while maintaining the capability to represent an original design concept with combinations of the images. The size of the mood board was also considered in terms of the participants' experiment logistics and task load as we planned to conduct a large-scale experiment. The MBC uses a DCP as queries and searches images from Adobe Behance(Wilber et al. 2017). The UI renders the upper right quadrant of the Character Space (CS), consisting of word 1 and word 2 as attributions on the semantic axes (Fig.1-C,D). The proposed MBC system is designed to encourage users to iterate the exploration of images till they are satisfied with the overall mood board composition. Various cost factors can hinder these iterative processes, such as the time and effort to collect materials, trying different search queries and re-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

Image References:

https://www.behance.net/gallery/113437545/CSW/modules/648144727

https://www.behance.net/gallery/82167757/COSH-series-of-armchairs-for-Boliacom/modules/476202235

https://www.behance.net/gallery/90111067/The-Hoffmann-Collection/modules/520974885

https://www.behance.net/gallery/145381227/Kioto/modules/821319521

https://www.behance.net/gallery/114951215/Wesley-Faux-Leather-Lounge-Chair-Set-3D-Model/modules/656201147

https://www.behance.net/gallery/13899989/Re-turned/modules/96236775

https://www.behance.net/gallery/149515177/Silky-(2022)/modules/844396881

https://www.behance.net/gallery/147913281/Dedas-public-seating-system/modules/835536071

https://www.behance.net/gallery/114951215/Wesley-Faux-Leather-Lounge-Chair-Set-3D-Model/modules/656201147

Figure 1: The proposed MBC system UI. It allows users to move any image within the 3×3 matrix. Moving images upward (G) will weight more of the semantics of the word 1 (C), "ergonomic", and moving images to the right (H) will weight more of the semantics of the word 2 (D)"comfortable" when it performs the next search.

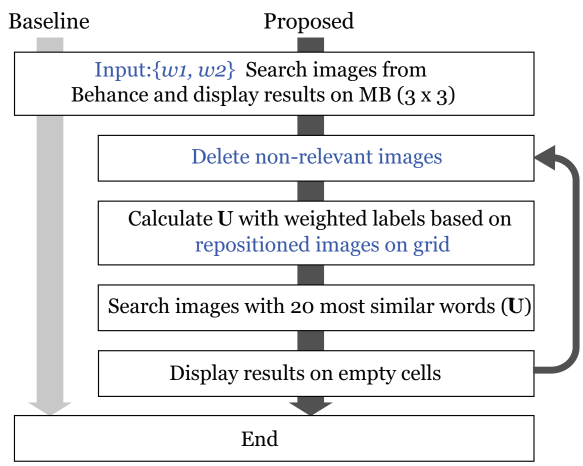


Figure 2: Variations of Mood Board Composer (MBC) algorithms. Text in blue shows the user's operations

sources, and figuring out the compositions. Edwards et al. (2009) discussed that the iterative process could be discouraged due to the vast choice of images offered by digital resources. They further argued that once images are selected, confidence is built so that users feel the continuous search for new material would be futile. To confirm the positive effect of iterative processes and overcome this iteration cost, we designed our experiment in the following ways. We first set up a comparative experiment between a baseline system that does not involve users' iterations for composing a mood board and the proposed system, which allows users to iterate as many as they like. We aimed to implement low-cost and high-engagement interaction so that the users can effortlessly try the optimum number of iterations to get the best experience in mood board creation tasks. We developed two separate systems, the baseline system, which does not support iterations, and the proposed system. Figure 2 shows the overall differences in the algorithms each system takes.

Baseline Search Algorithm

The baseline tool first receives the user's query (Q) input as two adjectives, (w_1, w_2), in the two search windows (Fig.1-A). When the "START" button is pressed, the system will search images on Behance in three "Fields," which are "Industrial Design," "Architecture," and "Fashion," which are likely to contain more of style elements such as form and CMF (Color, Material, and Finish) than, for instance, "web and graphic design" does. The candidates of images are ranked by relevancy and sorted per field. The top nine images are then randomly assigned to an empty grid of the 3×3 image set (D) of the mood board. This single session concludes the algorithm, and the user can export the mood board as a PNG file.

Proposed algorithm - Query update with average vector calculation

The proposed algorithm (Algorithm 1) involves query modifications based on user feedback. For each image on the current mood board, the system acquires semantic labels from the Google Vision API (Chen and Chen 2017). The Vision API uses pre-trained machine learning models, assigns labels to images, and classifies them into millions of predefined categories. The proposed system obtains the top five labels for each image on the mood board, ranked by the confidence score. Let $D\,(d_1, d_2, ..., d_m)$ be the image set on the current mood board, where d_i is the i-th image on the mood board, $L^i\,(l^i_1, l^i_2, ...l^i_k)$ be the labels for each image, where l^i_j is the j-th label for image d^i, and $S^i\,(s^i_1, s^i_2, ..., s^i_k)$ be the confidence score from the Vision API assigned to each label, where s^i_j is the score of the j-th label for image d^i. For each image label l^i_j in the set of image labels L^i nested under each image d_i on the mood board D, the system assigns label vectors using Concept Net Numberbatch word embedding. Let $\boldsymbol{V}^i\ (\boldsymbol{v}^i_1, \boldsymbol{v}^i_2, ..., \boldsymbol{v}^i_k)$ be the vectors of the labels $L^i\ (l^i_1, l^i_2, ...l^i_k)$, where $\boldsymbol{v}^i_j$ is the vector of the j-th label for image d^i. A mean vector $\overline{\mathbf{v}}_i$ of the image d_i can be calculated as follows:

$$\begin{aligned} \overline{\mathbf{v}}_i &= \frac{(s^i_1 \boldsymbol{v}^i_1 + s^i_2 \boldsymbol{v}^i_2 +, ..., s^i_k \boldsymbol{v}^i_k)}{k} \\ &= \frac{1}{k}\sum_{j=1}^{k}\{s^i_j \boldsymbol{v}^i_j\} \end{aligned} \tag{1}$$

where k is the total number of labels for image d_i.

Algorithm 1 Proposed (updating query)

1: **function** NEWQUERY
2: $Q^{new} := []$
3: $L, S, V := []$
4: $Wt := []$
5: $\overline{\mathbf{v}}_i := []$, ${}_{Weighted}\overline{\mathbf{v}}_i := []$
6: $\mathbf{U} := []$
7:
8: **for** each d_i in D **do**
9: L.append(VisionAPI(d_i))
10: S.append (VisionAPI(d_i))
11: Wt.append (OnDropWeight(x, y))
12: **for** each l_i in L **do**
13: V.append (ConceptNetVector(l_i, s_i))
14: **if** $cosSim(l^i, w_1) > cosSim(l^i, w_2)$, **then**
15: ${}_{Weighted}\overline{\mathbf{v}}_i = \overline{\mathbf{v}}_i \times Wt(\beta)$
16: **else**
17: ${}_{Weighted}\overline{\mathbf{v}}_i = \overline{\mathbf{v}}_i \times Wt(\alpha)$
18: **end if**
19: **end for**
20: $\mathbf{U}$:= Mean (${}_{Weighted}\overline{\mathbf{v}}^i$)
21: **end for**
22: Q^{new}.append (**MostSimilarWords** ($\mathbf{U}$))
23: **end function**

The proposed system lets users reposition images on the mood board's 3×3 matrix. This operation determines which of the labels on images should be enhanced towards the semantics of either word 1 or word 2 by classifying the image labels into two classes, w_1_labels, and w_2_labels. Then, only one of the pairs of position weights, $Wt(\alpha, \beta)$ (Fig.3), assigned to each grid is multiplied for the labels that are classified as the class of label. This classification is performed by comparing the cosine similarity ($CosSim$) of each label to the vector of w_1 and w_2 (Algorithm 1-14). For example, if a label vector is more similar to the meaning of w_1, the label is classified as a w_1 label, and the label vector is multiplied only by the β value (w_1 on y axis side) of the pair of position weight $Wt(\alpha, \beta)$. This way, the user's repositioning an image towards a particular direction on the matrix will provide feedback to the system (Fig.1-G,H). The system, in effect, will detect the users' intention to enhance a particular semantics in the following search without having to modify the query explicitly. The position-weighted average vector ${}_{Weighted}\overline{\boldsymbol{v}^i}$ of the repositioned image d_i will be updated as described in Algorithm 1 (14 -17). As for the paired weight for each position in the 3×3 grid, which will be multiplied by a label vector, we have tested two options with several initial queries. Figure 4 shows the weight array we implemented. It keeps the images fairly close to the user's intention while expanding the semantic space to explore.

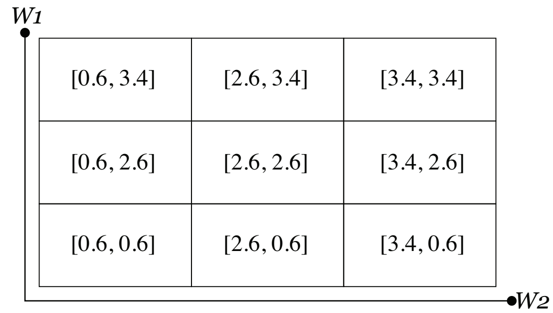


Figure 3: Pairs of position weights $Wt(\alpha, \beta)$ on the mood board matrix. These weights are assigned upon dropping the image to (x, y) coordinates.

The final step before updating the new query is to get the average vector of all the current images on the board, which can be calculated as follows. Let $\mathbf{U}$ be the average of all the weighted vectors for the images $\{{}_{Weighted}\overline{\mathbf{v}}_1, {}_{Weighted}\overline{\mathbf{v}}_2, ..., {}_{Weighted}\overline{\mathbf{v}}_m\}$ on the board.

$$\mathbf{U} = \frac{1}{m}\sum_{i=1}^{m} {}_{Weighted}\overline{\mathbf{v}}_i \tag{2}$$

where m is the number of images on the current mood board.

Calculating most similar words To update the query for the next search, the system will get the top 20 most similar words according to the input, in this case, $\mathbf{U}$, the average vectors of all weighted vectors for images on the mood board D. The system computes the cosine similarity ($CosSim$) with the normalized input vectors and outputs the top-N words in $CosSim$. This function is implemented as a method in a Python package, gensim.models(Srinivasa-Desikan 2018).

Experiment Design

The study protocols below have been approved by the Institutional Review Board of the National Institute of Informatics, Tokyo, Japan (Approval number 0042).

Participants and Independent Variables

120 participants, whose job function was "Arts, Design, or Entertainment and Recreation" and who was fluent in English, were recruited via Prolific(Palan and Schitter 2018). 11(9.17%) did not complete the study due to system trouble or unknown reasons. This left us with a total of 109 participants (50 M, 55 F, 4 Non-binary) who completed the study, with a mean age of 33.00 years (σ = 11.34). The participants who completed the study were paid US$12. All

participants were asked to perform the mood board creation task twice with the same type of MBC system. The between-participant factor was the difference in the used MBC system (Fig.2), and the within-participant factor was the two different Design Concept Phrases (DCP) they were given to use as the initial query Q. The factor incorporated in these two DCPs was the $CosSim$ between word 1 and word 2 in the DCP. The near DCP was "Ergonomic Comfortable," and the far DCP was "Relaxed Skillful." The $CosSim$s of those two DCPs were 0.4528 and 0.0053, respectively. The order of the DCP they used in the two tasks was assigned randomly in a counterbalanced order.

Dependent Variables

We used the Creativity Support Index (CSI)(Cherry and Latulipe 2014) as a post-task psychometric measurement to compare four conditions, with a baseline MBC and an experiment MBCs, in terms of supporting creativity in a mood board composition task. The CSI The CSI has a rigorous protocol, which evaluates the result of creation in relation to a user's effort, such as "I was satisfied with what I got out of the system or tool." This is suitable for tools designed for experienced users who know what the creative outcomes are and what the ideal experiences in creation are.

In addition, we employed a single-item measurement for remaining mental resources, the Gas Tank Questionnaire (GTQ)(Monfort et al. 2018), immediately before and after each task. The GTQ attempts to measure users' cognitive load who engage in a task without burdening them by asking multiple questions. The GTQ asks a question, "Think about your brain as an engine. Slide the fuel tank indicator (0 to 100) below to show how much gas you have left now." We took the differences between Gas Tanks before and after the task as a value that indicated the mental resources consumed to perform the task.

Stimuli and Tasks

Participants were randomly assigned to either of the four groups, two counter-balanced groups in different distance DCPs for each baseline and experiment tools and given instructions on the experiment. The MBC tool was provided to the participant as a web link along with the DCP. The participants were asked to download the mood boards they created to their local computers and upload them to the questionnaire on Survey Monkey. They then went through all the CSI questionnaires, followed by the second pre-task GTQ, the second task with the same tool and the other DCP, and the second post-task GTQ. Finally, they responded to a Paired-Factor Comparison that gave weight to each category of CSI evaluations across both tasks.

Results and discussion

Note that of all the 218 cases, 4 cases were disqualified because their responses to the CSI questionnaire had identical scores all the way through the survey (all 0 or all 10), and 3 cases were excluded because they did not upload valid mood boards, which left us with 211 cases (58 baseline and 49 proposed tool cases) for the final analysis.

Table 1: Mean CSI and Mental Load (GTQ) by tool

	Baseline(σ)	Proposed(σ)	p	Cohen's d
CSI	$21.05(12.36)$	$52.57(24.98)$	$< 0.001^{**}$	1.64
GTQ	$-0.86(9.89)$	$2.85(8.74)$	0.04^{*}	0.40

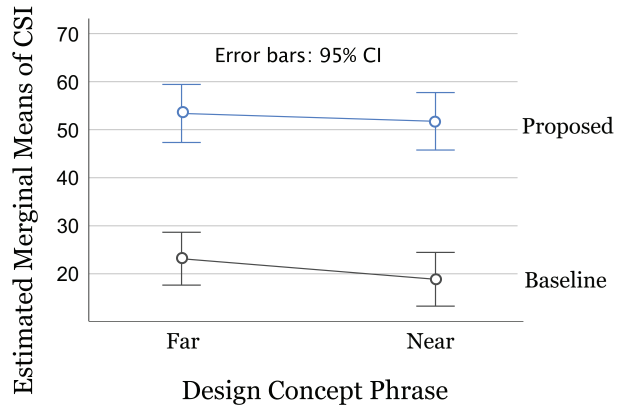


Figure 4: Variance in mean CSI for two different DCPs.

Creativity Support Index

Table 1 shows the mean CSI score and the mental road, measured by the GTQ, between by the tool. The CSI score of the proposed tool was significantly higher than that of the beseline tool, and the mental load of the proposed tool was significantly higher than that of the baseline tool.

Within Participant Factor

Fig. 4 shows the variance in the tool's estimated marginal means of the CSI scores for two different DCPs. The CSI scores with the baseline tool with far and near DCP were 23.15 and 18.87, respectively, which had a significant tendency ($p = 0.064$). With the experiment tool with far and near DCP were 53.40 and 51.76, respectively, which was not significant.

The proposed algorithm, which allowed the participants to iterate the image search interactively, was valid in supporting creativity in the mood board composition task, demonstrated by the CSI score. The values of the pre-task and post-task GTQ between the proposed tools and the baseline tool suggested that the users may have felt exhausted by the operation they had to follow on the proposed algorithm. However, the CSI score clearly shows that the cost is worthwhile. On the other hand, the proposed tool may have left users unclear about how repositioning images on the grid exactly works. Meanwhile, the CSI score difference between far DCP and near DCP seemed to be more apparent with the baseline tool than with the proposed tool. This implies that a potential disadvantage of near DCP may have been compensated by the proposed tool when users are engaged in a visual task.

One limitation of the work is that the way we set up the experiment in comparing the effectiveness of the proposed algorithm to the baseline tool. While we did not find comparable prior studies which uses the grid system to compose a mood board, we had to rely on a rather an artificial baseline tool on our own. In the future we plan to compare variations of iteration algorithms to compare what element of the iterative algorithms, for example, comparing repositioning the images on the board vs. operating the semantic labels on each image directly, and so on. Also, more detailed analysis on the factor scores of the CSI may reveal which aspect of the creativity was supported by what algorithms, which is also our future work.

Conclusion

Through experimenting with the two different MBC tools, we confirmed the effectiveness of the iterative process that allows user feedback, making the mood board creation task more engaging for concept designers. The present study contributes to the field of computational creativity by offering adaptive query updates utilizing the 2-D semantic space where users can rearrange the images on the mood board. Our post-hoc analysis of the CSI and GTQ scores suggest the participants may have been exhausted by the complex process of iterations, yet the effect of the creativity support overcame the cost. We also observed that the characteristics of the initial verbal query may be a strong factor for users to feel creative about the concepts they are operating, but the proposed tool may close such gaps.

References

Aliakseyeu, D.; Subramanian, S.; Lucero, A.; and Gutwin, C. 2006. Interacting with piles of artifacts on digital tables. In *Proceedings of the working conference on Advanced visual interfaces*, 159–162.

Bouchard, C.; Camous, R.; and Aoussat, A. 2005. Nature and role of intermediate representations (ir) in the design process: Case studies in car design. *International journal of vehicle design* 38(1):1–25.

Cassidy, T. 2011. The mood board process modeled and understood as a qualitative design research tool. *Fashion Practice* 3(2):225–251.

Chen, S.-H., and Chen, Y.-H. 2017. A content-based image retrieval method based on the google cloud vision api and wordnet. In *Asian conference on intelligent information and database systems*, 651–662. Springer.

Cherry, E., and Latulipe, C. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21(4):1–25.

Edwards, A.; Fadzli, S.; and Setchi, R. 2009. Comparative study of developing physical and digital mood boards. In *Anais... 5th International Conference on Innovative Production Machines and Systems. Carfiff, UK.*

Koch, J.; Lucero, A.; Hegemann, L.; and Oulasvirta, A. 2019. May AI? design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Koch, J.; Taffin, N.; Lucero, A.; and Mackay, W. E. 2020. Semantic collage: Enriching digital mood board design with semantic labels. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 407–418. Association for Computing Machinery.

Li, T.-t., and Zhao, D.-h. 2021. Mood board tool on high-level semantics visual representation to favor creative design. In *International Conference on Applied Human Factors and Ergonomics*, 408–419. Springer.

Lucero, A. 2012. Framing, aligning, paradoxing, abstracting, and directing: how design mood boards work. In *Proceedings of the designing interactive systems conference*, 438–447.

Monfort, S. S.; Graybeal, J. J.; Harwood, A. E.; McKnight, P. E.; and Shaw, T. H. 2018. A single-item assessment for remaining mental resources: development and validation of the gas tank questionnaire (GTQ). *Theoretical Issues in Ergonomics Science* 19(5):530–552.

Palan, S., and Schitter, C. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17:22–27.

Sano, S., and Yamada, S. 2022. Ai-assisted design concept exploration through character space construction. *Frontiers in Psychology* 12:6722.

Setchi, R.; Tang, Q.; and Stankov, I. 2011. Semantic-based information retrieval in support of concept design. *Advanced Engineering Informatics* 25(2):131–146.

Srinivasa-Desikan, B. 2018. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

Wilber, M. J.; Fang, C.; Jin, H.; Hertzmann, A.; Collomosse, J.; and Belongie, S. 2017. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, 1202–1211.

Zabotto, C. N.; Amaral, D. C.; Hornos, C. J. M.; Benze, B. G.; et al. 2019. Automatic digital mood boards to connect users and designers with kansei engineering. *International Journal of Industrial Ergonomics* 74:102829.

The Lena Singer Project: Simulating the Learning Experience of a Singer

Matthew Rice[1] and **Simon Colton**[1,2]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]SensiLab, Faculty of Information Technology, Monash University, Australia

m.rice@se22.qmul.ac.uk s.colton@qmul.ac.uk

Abstract

The Lena Singer project involves a generative AI process based on a recent singing voice synthesis system that iteratively produces audio to simulate a singer learning how to sing. Users can select an initial motivation and an initial ability for the singer, then, through a feedback-based process involving random elements, the singer may improve at singing, or they may get worse, which in turn boosts or diminishes its confidence, ability, and motivation. In this way, we aim to provide a simple model which simulates the learning experience of a human singer and demonstrate how it differs from standard machine learning approaches. We also explore the feedback loop that learning can have on internalized features, and exemplify how a machine might express the output of this learning. Finally, we discuss how the context provided by this process can be seen as relatable. A demo of this project can be found at `https://lena-singer.vercel.app`.

Introduction and Background

Since the introduction of artificial neural networks, the goal has been to emulate human learning by modeling the brain and the composition of its neural connections (Russell and Norvig 2010). However, even current deep learning or reinforcement learning systems can't model some of the complex ways particular factors or experiences can affect decision-making and learning (Simplilearn 2022). Moreover, emulation systems are designed to approximate some human capabilities, but there is a stark difference in how these systems are trained (by minimizing some objective functions) versus how a human would learn through various experiences and situations. This is particularly seen in the fundamental differences in biological and artificial intelligence (Korteling et al. 2021).

When people learn, important factors related to self-theories, i.e., people's beliefs about themselves, can have a huge impact. In the influential human psychology book "Self-theories: Their Role in Motivation, Personality, and Development", Dweck frames students' response to failure into two categories: 'helpless' and 'mastery-oriented' (Dweck 2000). The students in the helpless category, although at the same initial ability as the others, tended to lose motivation and give up, whereas the mastery-oriented students would tend to be resilient. However, sorting into these two categories seems to directly relate to the students' learning goals, which are determined before they even start learning. Other research in (Druckman and Bjork 1994) suggests self-confidence is related to one's perception of ability, and may play a central role in how one learns skills over time. Moreover, these variables seem to be part of a feedback loop, as confidence can increase motivation, which in turn can increase ability and this can cyclically increase confidence (Bénabou and Tirole 2005).

We aim here to simulate some of these internal factors and demonstrate how their initial state may play a substantial role in an agent's ability to learn a creative skill such as singing. Learning to sing well can take years, and often involves thousands of hours of repetitive performance and analysis of one's own singing voice as it would compare to others. This process can require a great deal of patience and continued motivation to keep practicing. Moreover, a lack of belief in oneself, due to low confidence, is a common cause of mistakes in singing (Ni Riada 2019).

Recent advances in generative deep learning have produced high-quality singing voice synthesis (SVS) systems. These systems can generate audio of a realistic human singing voice from musical scores and lyrics. In this project, we use VISinger2 (Zhang et al. 2022), a recent high-quality SVS system. We combine this with a step-based probabilistic learning model that we heuristically construct to converge to particular outcomes. The outputs of the probabilistic learning model are used to corrupt the inputs and outputs of the singing model to alter the singer's ability at each step. Via a web-based front end, users can define the starting state of the model and watch as it either improves at singing or ultimately performs worse. Our contributions include:

1. The development of a novel initial framework for simulating aspects of human learning, built on top of an existing controllable AI generation system.
2. The exploration of how a machine can make mistakes and how these may differ from those made by people.
3. A demonstration of how the learning process of a relatable creative AI system may reassure users and encourage them to build a narrative about the process.

ISBN: 978-989-54160-5-9

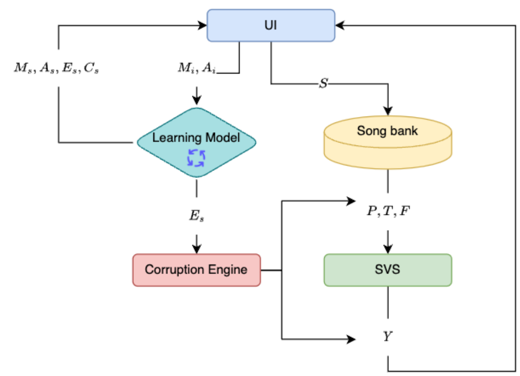


Figure 1: Overview of the Lena Singer system which shows the major components and data flow. P, T, F represent phonemes, phoneme timings, and frequencies respectively. M, A, E, C represent motivation, ability, mistakes, and confidence. Y represents the output audio.

System Description

The Lena Singer system is composed of four parts: the *singing model*, the *learning model*, the *corruption engine*, and the *user interface*. Figure 1 presents a diagram of how the component parts work together in the overall system.

Singing Model

The underlying SVS system is VISinger2 (Zhang et al. 2022), an end-to-end model that creates a realistic singing voice in Mandarin Chinese from an input of phonemes, phoneme timings, and phoneme frequencies. At its core, the model is a conditional variational autoencoder with a discriminator. The model encodes phonemes, phoneme timings, and frequencies to a latent space and similarly encodes the mel spectrogram (O'Shaughnessy 2000) of an associated audio file. During training, the objective is to minimize the distance between the latent encodings of related phonemes and spectrogram representations. The model also includes a decoder, which uses a DDSP (Engel et al. 2020)-inspired process to generate a mel spectrogram, which is converted to the waveform domain using a modified HiFiGan (Kong, Kim, and Bae 2020) model. Following the decoder, a discriminator is simultaneously trained to ensure high-quality outputs that are similar to the training examples.

The developers of VISinger2 used the Opencpop (Wang et al. 2022) dataset to train the model. This consists of 100 Mandarin songs sung by a professional singer with human-labeled annotations. The annotations include the lyrics, the phonemes, the notes, the note durations, the phoneme durations, and whether or not a note was a slur note. We obtained permission to use the Opencpop dataset and trained our own version of the model accordingly, using a single A100 GPU, for 200k steps. The fine-grained control of phonemes and pitches allows for precise and realistic manipulation of the resulting audio. As part of the Lena Singer system, this SVS system can generate one of five songs from the Opencpop test set, with each song being around seven seconds long. Importantly, the model is used purely to generate audio from the inputs given to it by the corruption engine and its outputs are not fed back into the learning model. Thus it could be easily swapped for any other SVS system and/or dataset providing it follows the same format as the Opencpop dataset.

Learning model

In overview, we aim to implement a system able to simulate the human learning process, using variables that represent relatable concepts. Specifically, we develop a simple stochastic model for how factors of motivation, confidence, ability, and mistakes during performance interplay over time, as an agent practices the creative act of singing. Overall, the learning model consists of seven variables with a range of 0-100 which are updated at every step, based on the values of the other variables. They represent both the internal state of the singer (ability, motivation, confidence, mistake factor, mistake history factor) and the external results (mistakes/mistake history). All variables are updated in a stepwise fashion and are normally distributed with a standard deviation of 5. This allows for some randomness, hence different results from the same starting point. Algorithm 1 shows the pseudocode for the variable updating scheme. We designed this model with three goals in mind:

1. To simulate how a large amount of initial motivation or ability should usually be able to overcome a low amount of the other. Low amounts of both variables should usually lead to failure, while high amounts should usually lead to success.
2. To ensure that the system is non-deterministic, so two sessions with the same initial values of ability and motivation may lead to different processes and outcomes.
3. To simulate how learning to sing well should take more timesteps than giving up and stopping.

The variables in the learning model are defined as follows: **Ability**: the singer's natural ability. It is inversely related to mistakes but also affected by current motivation. **Motivation**: The singer's interest in continuing to learn. A low motivation will cause the singer to stop trying to learn and give up. **Confidence**: how confident the singer is in its abilities. A high confidence will cause the singer to stop trying to learn because it believes it is good enough. **Mistakes**: how corrupted the output audio will be. **Mistake Factor**: how much the current mistakes value matters to the singer. **Mistake History**: how many past mistakes the singer remembers. **Mistake History Factor**: how much these past mistakes matter to the singer. The mistake factor, mistake history, and mistake history factor variables were designed to allow the model to learn to overcome and ignore their current and previous mistakes as a representation of resiliency.

Although the variables and update equations are not explicitly derived from any true academic model, they were initially inspired by human learning, then fine-tuned and weighted to fit our goals. They rely on a feedback model of learning where internal state variables are updated from combinations of other internal variables and external results.

Algorithm 1 Updating scheme for the learning model

$let\ N(x) = \mathcal{N}(x,\, 5)$
motivation $m \leftarrow m_{init}$
ability $a \leftarrow a_{init}$
confidence $c \leftarrow 0$
mistakes $e \leftarrow 0$
mistake factor $mf \leftarrow 0$
mistake history $h \leftarrow []$
mistake history factor $hf \leftarrow 0$
step $n \leftarrow 0$
while $n < 30$ **do**
  $e \leftarrow N(100 - 0.55a - 0.45m)$
  $h.append(e)$
  $h.resize(10 - c/10)$
  $c \leftarrow N(a - e * mf)$
  $a \leftarrow N(100 - e + m * n)$
  $m \leftarrow N(a - \sum h * hf)$
  $mf \leftarrow m * 0.01$
  $hf \leftarrow (100 - m/100) * 0.01$
  $n \leftarrow n + 1$
  if $c = 100$ *or* $m = 0$ **then**
    break
  end if
end while=0

For instance, confidence is derived from the agent's current ability as well as current mistakes multiplied by a mistake factor. Over a maximum of 30 steps, the model will converge to one of three outcomes. If confidence continues to stay near 100, the model will stop to signify that it has reached its goal. Or, if motivation hovers near 0, the model will stop to show it is giving up. Finally, if the model reaches the maximum number of steps, the model will stop to denote that it is finished learning. The model itself depends solely on these variables and does not rely on feedback from the corruption engine, SVS, or the UI.

Corruption Engine

The corruption engine produces the mistakes the singer is making. There are nine corruptions split into pre- and post-generation corruptions. The pre-generation corruptions affect the inputs to the VISinger2 model. These include overall speed change, inter-phoneme timing change, and random phoneme replacement. By reducing or increasing the phoneme timings, either by a global amount or a varying amount, the speed of the singing is affected without changing the pitch. The post-generation corruptions are high and low-pass filtering, compression, distortion, random pitch detune, and gain reduction. Besides random pitch detune, which is implemented directly, all other effects are implemented using Pedalboard (Sobot 2021). There is also a reverb effect added that inversely follows mistakes to frame the perfected singer as more professional. Uniform distributions are used to regulate each corruption's activation and intensity, which are parameterized by the mistakes metric or, in the case of gain reduction, the confidence metric. The ranges for these distributions are chosen heuristically.

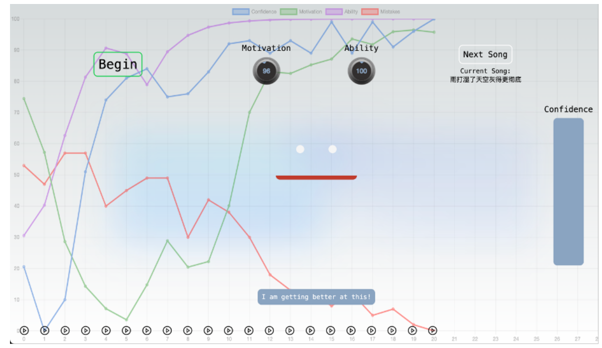


Figure 2: The web interface for the Lena Singer system. This run had an initial motivation of 74 and an initial ability of 31. Here, the agent successfully learned how to sing.

User Interface

The GUI of the system enables users to set the initial motivation, ability, and song and displays the output data from the learning model and SVS model. The motivation, ability, confidence, and mistakes metrics are plotted on a graph for each time step, while the audio output appears at the bottom of the screen as a button. The GUI will also display some text to reflect the current state of the learning run. Figure 2 shows an example session in the GUI.

Experimental Results

We evaluate here both our learning model and the corruption engine. For all evaluations, we used a default song, as song choice has no impact on the learning model. To evaluate the learning model, we ran 200 simulations of the model with random initial states of motivation and ability, sampled on a uniform distribution from 0-100. Figure 3 shows if the run was a success as well as how long it took based on the initial variables of ability and motivation. Overall, it seems like our initial goals for the learning model have been achieved in terms of sensible outcomes from certain starting conditions. In particular, a high ability or high motivation will likely lead to success, while mediocre or poor metrics lead to failure. Also, it seems there are many cases with a similar initial state that lead to different results. Furthermore, in general, it seems like achieving the task of singing takes longer than failure. However, it is interesting that having a high initial ability and motivation does not seem to generally converge to success quicker than in other cases.

As a straightforward way to evaluate the corruption engine, we measure the multi-resolution STFT (Steinmetz and Reiss 2020) error between the clean audio (no corruption) and the output audio at every time step for a particular run with initial motivation = 86 and initial ability = 25. Figure 4 plots these metrics. It seems that the mistakes graph somewhat correlates with the multi-resolution STFT error (PCC of 0.778). Therefore, the system reduces the quality to some extent depending on the mistakes metric. However, we argue that there is no requirement for a perfect linear relationship in this emulation system.

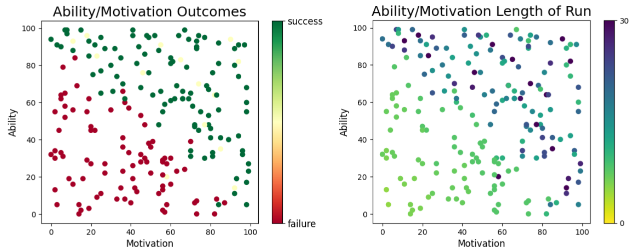


Figure 3: Success/failure and number of steps taken for 200 initial states of motivation and ability.

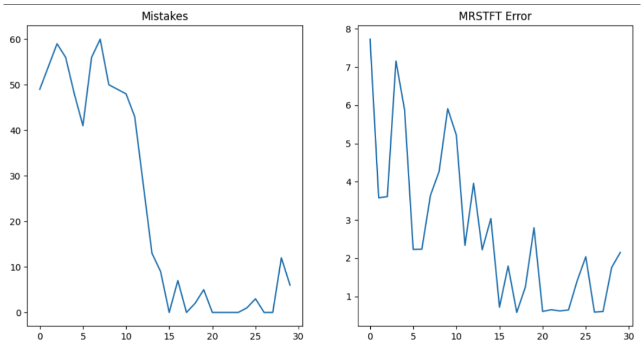


Figure 4: Mistakes and MRSTFT error between corrupted audio and clean audio on each learner model step.

Discussion

Although this project represents a toy example of human learning, it achieves a result that is empirically much closer and more relatable to a user's actual experience learning a skill than standard machine learning approaches. Each simulation run produces a unique result in terms of the learner model variable journeys, as well as the output audio on each step. When running a simulation with the system, a user can experience and empathize with the learning agent, potentially building up a mental story as to why a particular outcome might have occurred. For instance, a run that begins with high confidence and high motivation, but high mistakes, may see the confidence and motivation drop as the errors increase, but then slowly start to decrease the mistakes, leading to an increase in confidence. This could cause the user to imagine a singer initially frustrated with their mistakes, but with enough resilience to overcome them and gain confidence. These runs may cause the user to feel more connected to both the intermediate and final outputs. However, as suggested by (Colton, Pease, and Saunders 2018), barriers exist to truly make these "life experiences" challenging for people to accept, unless we are able to more accurately reflect the audience or alternatively use an enactive AI (Froese and Ziemke 2009) approach, with a completely different social and cultural environment.

While the learning model variables and update scheme were chosen arbitrarily, they seem to control the system such that the results are mostly expected, but sometimes surprising. The variables also present a feedback loop, as an increase in one usually leads to an increase in another. Although the learning model itself is significantly more interpretable than a standard deep learning approach, it is still not entirely clear how variable updates can cause particular outcomes. The corruptions selected for this system were primarily chosen for their ease of implementation. However, their outputs are significantly different from the consequences of human errors. While people cannot physically manipulate their voices with audio effects, some corruptions generate unnatural outputs. Nonetheless, there are some similarities between the two, e.g., it is common for a human singer to rush through a song or sing more quietly due to nervousness or lack of confidence. We didn't strive to completely model human mistakes but attempt to illustrate some initial examples to inspire future discussion.

Related Work

To the best of our knowledge, modeling motivation, ability, and confidence in a learning model for a creative skill like singing hasn't been studied from a computational creativity perspective. However, there is related work on the creative process and intrinsic motivation, e.g., (Salge, Glackin, and Polani 2014) explores intrinsic motivation in AI, and builds a 3D simulation to explore a mathematical definition of agent empowerment as an example. In addition, (Guckelsberger, Salge, and Colton 2017) describes an enactive framework to map an AI's intrinsic values and goals to its creativity. Unlike our system, they develop a non-anthropocentric model and aim to study creativity from the bottom-up.

Earlier work on music generation with creative agents in (Miranda 2003) explored granular synthesis through imitation agents. However, the agents themselves don't have any sense of embodiment or self-awareness. (Linkola et al. 2017) focus on self-awareness as it relates to metacreativity and "the capability to reflect on one's own creative processes and adjust them". Their model defines key aspects of self-awareness that are useful even for non-metacreative systems, namely artifact-awareness and goal-awareness. (Ford and Bryan-Kinns 2023) study the aspect of reflection in people using creativity support tools, and suggest that it is an important part of self-expression. Finally, (Cook et al. 2019) discuss the idea of framing in computational creativity, defined as "providing a narrative context for the actions and motivation of the software". They conclude that projects that include descriptions of the underlying processes can help audiences to relate to them.

Conclusions and Further Work

We presented a reasonable computational simulation of people learning how to sing. In particular, we developed a novel learning model and data corruption engine that attempts to model particular aspects of human learning in a feedback loop. We combined those modules with a recent controllable SVS model to synthesize realistic human singing, using the corruption engine to modify the inputs and outputs of the SVS model to portray the human learning process. We discussed the design decisions for our system and showed how the system meets our intended goals through the evaluation of experimental results. Furthermore, we described how such a system could be seen as relatable to a user. This system is only a first step for future systems studying human and machine learning through the lens of creative practice.

In general, this framework itself needs refinement and iteration. As a first step, future systems could attempt to have a richer and more scientifically accurate model of human learning for creative tasks. For example, in the previously cited work (Dweck 2000), researchers found that, while confidence is a good predictor for academic achievement, it doesn't help students in difficult situations. Alternatively, a model could be developed that would be more in line with enactive AI with intrinsic motivation, with a clear design to allow the model to have intentional creative agency that adapts to its environment. Moreover, while the corruptions provide a solid baseline, future work could explore either the idea of closer modeling of human mistakes or could posit novel corruptions that non-anthropocentric creative agents could explore. We would also like to emphasize the conclusions of (Shneiderman 2020) which suggest that assumptions from tool-like application systems, such as virtual assistants, should not be directly applied to emulation systems like the Lena Singer system. These should be treated and designed separately to avoid poorly crafted designs since they usually have separate goals. Finally, due to the advent of AI systems that can realistically mimic particular human abilities like singing, painting, etc..., we are particularly interested in work that follows a similar framework with different underlying generative AI technologies.

Author Contributions

Matthew Rice theorized and developed the Lena Singer system and web demo. Simon Colton provided advice and direction for this work. Matthew Rice wrote the majority of the manuscript, while Simon Colton contributed to the writing and editing of the manuscript.

References

Bénabou, R., and Tirole, J. 2005. Self-Confidence And Personal Motivation. *Psychology, Rationality and Economic Behaviour: Challenging Standard Assumptions* 19 – 57.

Colton, S.; Pease, A.; and Saunders, R. 2018. Issues of Authenticity in Autonomously Creative Systems. *Proc. ICCC* 272–279.

Cook, M.; Colton, S.; Pease, A.; and Llano, M. T. 2019. Framing In Computational Creativity – A Survey And Taxonomy. *ICCC* 156–163.

Druckman, D., and Bjork, R. A. 1994. *Learning, Remembering, Believing: Enhancing Human Performance*. Washington, D.C.: National Academies Press.

Dweck, C. S. 2000. *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press.

Engel, J.; Hantrakul, L.; Gu, C.; and Roberts, A. 2020. DDSP: Differentiable Digital Signal Processing. arXiv:2001.04643.

Ford, C., and Bryan-Kinns, N. 2023. Towards a reflection in creative experience questionnaire. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery.

Froese, T., and Ziemke, T. 2009. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence* 173(3):466–500.

Guckelsberger, C.; Salge, C.; and Colton, S. 2017. Addressing the “Why?” in Computational Creativity: A Non-Anthropocentric, Minimal Model of Intentional Creative Agency. *Proceedings of the 8th International Conference on Computational Creativity (ICCC'17)*.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. arXiv:2010.05646.

Korteling, J. E. H.; van de Boer-Visschedijk, G. C.; Blankendaal, R. A. M.; Boonekamp, R. C.; and Eikelboom, A. R. 2021. Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence* 4.

Linkola, S.; Kantosalo, A.; Mannisto, T.; and Toivonen, H. 2017. Aspects of Self-awareness: An Anatomy of Metacreative Systems. *Proceedings of the 8th International Conference on Computational Creativity (ICCC'17)*.

Miranda, E. R. 2003. On the evolution of music in a society of self-taught digital creatures. *Digital Creativity* 14(1):29–42.

Ni Riada, A. 2019. 3 Singing Mistakes That Are Killing Your Confidence. https://www.confidenceinsinging.com.

O'Shaughnessy, D. 2000. *Speech Communications: Human and Machine*. Wiley.

Russell, S. J., and Norvig, P. 2010. *18.7 Artificial Neural Networks*. Prentice-Hall, 3rd edition. 727–737.

Salge, C.; Glackin, C.; and Polani, D. 2014. Changing the Environment Based on Empowerment as Intrinsic Motivation. *Entropy* 16(5):2789–2819.

Shneiderman, B. 2020. Design Lessons From AI's Two Grand Goals: Human Emulation and Useful Applications. *IEEE Transactions on Technology and Society* 1(2):73–82.

Simplilearn. 2022. Artificial Intelligence vs. Human Intelligence | Simplilearn. https://www.simplilearn.com/artificial-intelligence-vs-human-intelligence-article.

Sobot, P. 2021. Pedalboard. 10.5281/zenodo.7817839. https://github.com/spotify/pedalboard.

Steinmetz, C. J., and Reiss, J. D. 2020. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*.

Wang, Y.; Wang, X.; Zhu, P.; Wu, J.; Li, H.; Xue, H.; Zhang, Y.; Xie, L.; and Bi, M. 2022. Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis. http://arxiv.org/abs/2201.07429.

Zhang, Y.; Xue, H.; Li, H.; Xie, L.; Guo, T.; Zhang, R.; and Gong, C. 2022. VISinger 2: High-Fidelity End-to-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer. http://arxiv.org/abs/2211.02903.

Towards the Automatic Customisation of Editable Graphics

Daniel Lopes, João Correia, Penousal Machado
University of Coimbra, CISUC, DEI and LASI, University of Minho
{dfl,jncor,machado}@dei.uc.pt

Abstract

More and more, graphic designers need to deliver faster and cheaper. To help speed up the creative process, we propose a computational approach for automatically stylising graphics so that these can relate to given semantic concepts. More specifically, by automatically selecting transformation and styling properties that better relate to a list of provided keywords, according to *ConcepNet*. For testing the presented approach, 2D posters referring to different concepts were stylised using our method and evaluated through a user survey. Although the system can be further improved, the results suggest the viability of the approach to aid the automatic styling of concept-related graphics, such as posters.

Introduction

In creative fields such as Graphic Design (GD), finding disruptive visual solutions that attract people's attention is of the utmost importance, either to create communicative or artistic design artefacts, e.g. so the designs stand out over other posters on the streets or other book covers on store shelves. However, as the urgency for more effective designs grows and the GD area is increasingly democratised, graphic designers need to deliver faster and cheaper, which often leads to the adoption of trendy solutions and precludes the exploration of innovative visual solutions.

This paper proposes a computational approach for automatically styling graphics so that these can relate to given semantic concepts. More specifically, we propose using *ConceptNet* (Liu and Singh 2004) to assess the semantic similarity between given keywords (that must be set by the user to describe a given concept) and labels (not changeable by users) given beforehand to a set of mutation methods and their respective parameters. The resulting similarity values are then used to calculate the probability of each mutation method and respective parameters to be used for styling a given graphical item, e.g. given the keyword *sky* a given item may be more likely filled in *blue* in detriment to other colours. In other words, given *sky*, if the method *fillColour* is picked, the most probable parameter would be *blue*.

Although our approach might be generic enough to be used in different creative contexts, in this paper it was tested in the generation of posters, i.e. for transforming and styling a number of text boxes, images, and geometric shapes within 2D pages. Furthermore, to create a tool that could be easily integrated into a designer's workflow, the presented approach was implemented as an extension for *Adobe InDesign* — a broadly used desktop-publishing software for GD. Thus, designers can alternate between manually and automatically editing items without leaving *Adobe InDesign*.

As seeking highly subjective concepts could lead to highly subjective results, the tests focused mostly on keywords that might often be related to visual stereotypes (at least, in Western culture), e.g. in Western culture, one might find the concept *love* often related to the colour *red*. Furthermore, other parameters were explored to understand whether we could lead the system to either focus on a principal idea or exploit a more diverse range of related ideas. Lastly, the generated posters were evaluated by means of a user survey. The latter suggests that, although the system can be further improved, for the tested experimental conditions, the present approach can be viable to automatically style graphics so that these visually relate to given keywords.

Related Work

To aid the generation of concept-related graphics, there have been theoretical studies about the relationship between words and visual features, e.g. the relation between emotions and features such as colours, shapes, directions, curve sizes, or edge types (Collier 1996; Cavanaugh, MacInnis, and Weiss 2016; Rodrigues, Cardoso, and Machado 2019).

Concerning practical applications, relating words to colours seems to be often explored. One can pinpoint datasets relating words to colours (Heer and Stone 2012) and even automatic systems to do so, e.g. by extracting colour pallets (O'Donovan, Agarwala, and Hertzmann 2011) or by recolouring bitmap images automatically (Lin et al. 2013).

Besides colour-focused approaches, one can pinpoint ones, for example, for editing texture-like vector images according to one given adjective (Heath and Ventura 2016), or to generate vector sketches that illustrate keywords by using Evolutionary Computation (EC) (Cunha et al. 2020) or Machine Learning (ML) techniques (Ha and Eck 2017) to interpolate existing sketches or even by composing existing 3D models (Coyne and Sproat 2001). Zhao, Cao, and Lau (2018) presented a system to learn the most attention-catching zones in posters according to a given set of themes, such as *minimalist* or *romantic*. This system can be helpful,

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

e.g. in the generation of layouts according to those themes.

The most popular approaches for translating semantic concepts into graphics nowadays must be those using transformers and stable diffusion to generate realistic bitmap images from given text prompts (Ramesh et al. 2022; Radford et al. 2021). A shortcoming is the enormous data and computation requirements needed to implement such systems. Furthermore, bitmap images are often not suited for creating GD artefacts, since designers often need to edit or update information or create different variations of the generated artefacts in different formats and sizes. Moreover, such systems often generate pastiche results (Toivonen and Gross 2015) (i.e. variations of existing styles). In that sense, although such systems can be helpful, e.g. to generate illustrative images, we believe these are not yet well suited for co-creatively styling GD artefacts such as posters.

Approach

For aiding the design of creative artefacts, we propose a computational approach to automatically select mutation methods and respective parameters (referred to as visual assets) that can relate to a set of keywords (i.e. given a concept) defined by users.

To demonstrate the proposed approach, we tested it for styling GD posters. We implemented it as an extension for *Adobe InDesign* so manual and automatic editions/mutations can be done interactively, fostering the collaboration between the system and human designers.

To use this system, human designers must start by inserting desired items (i.e. text boxes, geometric shapes, or images) into *InDesign* pages. Then, manual editions may or may not be done. Whenever desired, the user may define a set of keywords descriptive of the concept of the work and click a button in the extension to automatically style a number of items according to the defined keywords, e.g. by more likely selecting given colours in detriment to others or by more likely rotating items other than skewing them.

That is accomplished by picking mutation methods and respective parameters according to their semantic relation to the concepts the user wants to seek. *ConceptNet* (Liu and Singh 2004) is used to calculate the relatedness (semantic similarity) between a set of keywords (defined by the user) and a set of labels assigned in advance to the available mutation methods and respective parameters. As an example, one can retrieve such similarity values using *ConceptNet* queries such as: *https://api.conceptnet.io/relatedness?node1=/c/en/beach&node2=/c/en/yellow*. In this first iteration, the labels were defined by our research team. In future developments, we aim to gather these by means of a user survey. 357 labels were used (a complete list can be found in the supplementary materials).

The proposed procedure goes as follows: (i) Using the user interface, one must type an intended set of keywords, e.g. *sun* and *beach*; (ii) The system will call *ConceptNet* to calculate similarity values between each label and each keyword; the returned values range from -1 to 1, standing to highly dissimilar to highly similar, respectively; however, the values are truncated to range from 0 to 1, i.e. 0 is considered the minimum possible similarity; (iii) Each label will be assigned with the respective maximum similarity value, e.g. if the label *yellow* is considered 0.135 similar to *sun* and 0.056 similar to *beach*, then *yellow* will be assigned with the value 0.135; also, for each edition method, each parameter will be assigned with the maximum value of its labels, and each edition method with the maximum value of its parameters (maximum values are always picked so the system acknowledges the most important assets, whether these relate to a keyword or another); (iv) The assigned values are then used as the probability of each method/parameter being automatically picked to edit a given page item.

We implemented methods to edit the following properties: the shape of the surrounding box of the items, their size, position and order (z-position), flipping and blending modes, opacity, fill colour and tint, stroke colour and tint, stroke weight, rotation, and the item's shearing angle. Concerning text boxes, also text size, text colour, justification, vertical alignment, letter spacing, and line height were considered.

For the experiments hereby presented, the edition process went as follows: (i) For each page item, all the available methods were iterated and each one could or not be picked to edit the respective item, according to their assigned probability; (ii) Whenever a method was picked, one of the available parameters was automatically chosen using a roulette approach, according to the respective probabilities to run; (iii) Each selected method and its respective parameters were then used to mutate the given item.

Furthermore, to understand whether we could lead the system to either focus on a principal idea (i.e. to the most related assets) or explore a more diverse range of related ideas (i.e. a wider range of assets), we created two variables to remap the assigned probabilities exponentially. That is, we set up an exponent to remap method probabilities and another to remap parameter probabilities. We refer to these as probability exponents. In practical terms, by increasing these exponents, high-related assets will be even more likely to be picked than less-related ones.

Experimental Setup

Experiments were conducted to determine whether meaningful visual solutions could emerge using the proposed approach. The experimental setup was started by manually creating a base poster in *InDesign* (see Figure 1). This poster should be as visually neutral as possible while containing a reasonable amount of items, so reasonable visual changes could occur and thus stronger conclusions could be drawn. Therefore, 7 black and white items were used: 1 central circle, 4 rectangles dividing the poster into 4 parts, and 2 text boxes, one at the top and another at the bottom. There was no particular reason to choose such a layout rather than seek a medium-complexity, balanced, and neutral composition. The text within text boxes is changed along with experiments to relate to the respective keywords.

For each experiment, 30 posters were automatically generated from the base poster. Each poster was submitted to the edition process once (see Section *Approach*). Among the

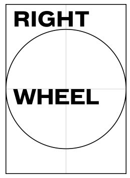


Figure 1: Base poster (created manually) from which the automatically-stylised posters were generated.

experiments, only the input concepts varied. As argued before, to avoid too much subjectivity, the tests focused mostly on concepts that we considered to have more or less obvious visual translations (at least in Western culture).

Moreover, to test the system in tasks of different difficulty, such concepts were selected (empirically) in a way that included keywords that were in two different categories: keywords that were equal to labels (similarity value equal to 1) and keywords that were not (similarity value smaller than 1). More specifically, we have selected 16 different keywords to build 10 different concept inputs: *"green", "big, love", "fear, dark", "sun, beach", "grape, vines", "small, shark", "big, shark", "baby, shark", "right, circle" and "colourful, balloons"* (an additional concept, removed due to space constraints, can be found in supplementary materials). 10 of those keywords were equal to labels of some method or parameter, i.e. *green, big, love, fear, dark, sun, small, right, circle,* and *colourful*.

For the aforementioned reason, one could expect the concepts *"green", "big, love", "fear, dark" and "right, circle"* to be the easiest to solve, *"baby shark" and "grape, vines"* the most difficult ones, and the remaining to be of medium difficulty. However, that might not always be true, as some mutation methods have a bigger perception impact than others, and the concepts refer to different levels of abstraction.

Although our approach can be used to pick any mutation method, to ease the interpretation of the experimental results, we have mainly focused on one of the most easily observable — colour. To do that, the methods to mutate fill colour and text colour have been set with a 100% chance of being picked regardless of the concept. The respective parameters (e.g. what colour to fill items with) and the remaining methods kept their automatically assigned probabilities. By experimentation, the method's exponent and the parameters' exponent were set to 3 and 5, respectively.

Experimental Results

Figure 2 showcases results for concepts in which keywords were used in the labelling of methods or parameters. Figure 3 showcases results for concepts in which one keyword (out of two) was included in labelling. Figure 4 showcases results for concepts in which none of the keywords was included in labelling.

The generated posters were evaluated by means of an online user survey (refer to supplementary materials). 34 people participated. To understand whether age, formation, or cultural background could influence the answers, such personal information was asked in the first instance. Also, visual disabilities were misguided. The respondents comprised an age group from 21 to 50, with 17 people being 21–22 years old. All were from Western countries (1 from Italy and 1 from the UK, both living in Portugal; the remaining were Portuguese), so one can assume all or most of them were familiar with similar cultural stereotypes. This was a relevant issue as, for different cultures, colours and symbols can have different meanings Since some questions referred to personal approaches to designing posters, the survey was specially directed to people with GD background. Even so, one of them had not. All questions were open-answer.

After gathering personal information, the respondents were asked what colours, shapes or other visual assets they would choose to design a poster for each of the keywords mentioned in Section *Approach*. The goal would be to later compare the assets chosen by the respondents with the assets automatically chosen using our approach. As previously indicated, the analysis of the answers focused mainly on colour. As the goal of our system is not to create composed or figurative images, such solutions were disregarded.

Colour-wise, the system matched the most mentioned choices of the respondents for each concept, i.e. green for "*green*", red and pink for "*big, love*", black and dark tones for "*fear, dark*", yellow and blue for "*sun, beach*", blue and grey for "*small shark*", "*big, shark*" and "*baby, shark*", colourful for "*colourful, balloons*", and purple and green for "*grape vines*". An exception is made for "*right circle*", for which the respondents more often chose the green colour and the system picked a variety of different colours. However, since the latter concept led to less consensual responses compared to the remaining, it could be predicted this concept would be more difficult to reach (same for an additional concept in supplementary materials: *right wheel*).

However, due to the set properties' exponent, the system often gave preference to one colour for each concept, whereas the respondents would sometimes use two colours in similar amounts, e.g. purple and green for *grape vines*.

Other than colour, the system sometimes matched the respondent's choices, at least for concepts included in labelling, e.g. it matched the circles in the *"right, circle"* posters, and the small items in the *"small, shark"* posters. However, further testing must be done to understand the system's effectiveness for methods that were not forced to always run, as *fillColour* and *textColour* were. Also, features that could not be achieved by the system were mentioned, e.g. repetition of items. In that regard, future work might comprise the development of additional mutation methods.

In question two, the respondents were asked whether the generated posters from Figure 2, 3 and 4 would better represent the respective concepts compared to the base poster, i.e. whether the system could improve the base poster conceptually. Also, they were asked to ignore legibility issues.

For 8 out of the 10 concepts (11 if counting with *"right, wheel"*), the majority of the respondents considered the posters improved. However, as could be predicted, for *right circle*, the respondents considered the posters did not represent the concept well. Suggestions included using circles rather than ellipses and adding more green.

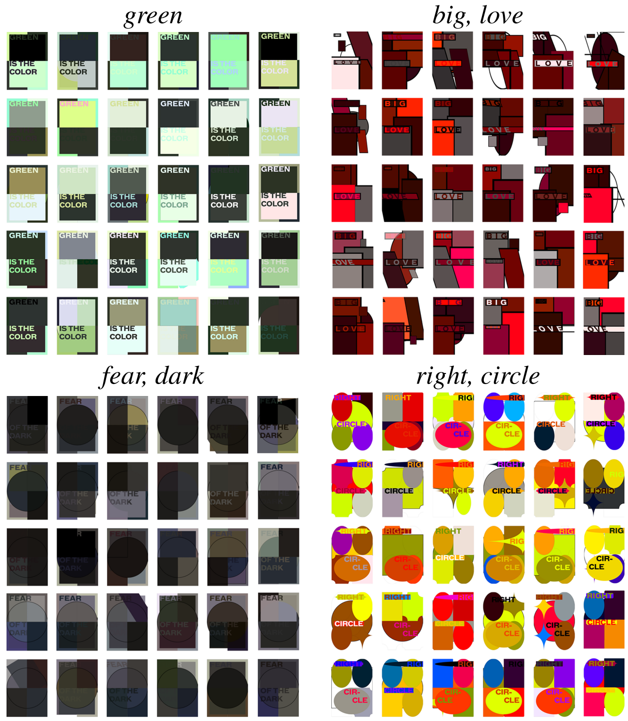


Figure 2: Posters generated from keywords used for labelling mutation methods or parameters.

Figure 3: Posters generated from one keyword used for labelling mutation methods or parameters, and one that was not.

Another non-consensual concept was *"small, shark"*, as 13 respondents considered there was no improvement, 11 that it did, and 9 that it did in part. The answers suggested that *small* was well represented. However, the rectangles did not seem to represent well the idea of *shark*, neither for *small*, *big* nor *baby, shark*. As the rectangles were already there in the base poster, it is likely that the system simply did not choose to make shape changes. Nonetheless, it would be positive if the system had transformed those rectangles into triangles, as frequently chosen by the respondents in the first question. For *"big, shark"*, adding red was also suggested. However, the big, no-fill rectangles with thick black strokes seemed to recall a more aggressive look, which pleased most respondents. Further suggestions went through decreasing dark tones for the *"green"* concept, using circles for *"grape, vines"* and *"colourful, balloons"*, and lighter colours for *"big, love"*.

In the third question, the respondents were asked whether they believed the generated aesthetics could already be used to create final GD posters, taking into consideration that the text contents could be changed. The answers suggest the respondents could see such potential in at least some of the posters. The most mentioned ones (by 20, 19, and 28 people, respectively) were the posters for *"fear, dark"*, *"colourful balloons"* and *"sun, beach"*. Furthermore, their comments suggest that illustrating given concepts can have significantly different objectives compared to designing a poster for communicating information about something related to that same concept. That can be noticed from the fact that many respondents seemed to be comfortable abdicating from the circle shapes (many times referred to in previous questions) if they were to design a poster for disseminating events related to grape vines or colourful balloons. Also, some respondents commented that the generated posters could be used as a design base, and small changes could then be manually made to create the final posters. A last noteworthy comment refers to the fact that some posters ended up being different from what the respondent chose in the first question. However, the generated solution still worked out well, suggesting the potential for the system to aid unexpected creative choices too.

The last question meant to understand whether a similar approach could aid creative (or not creative) choices in other areas rather than design. Furthermore, we asked which areas would those be. 25 respondents indicated that such approaches could be useful to hasten the beginning or during creative tasks, as long as they are used in a co-creative way, i.e. humans must be able to fine-tune the generated results. Besides some respondents still referred to GD applications, such as creating book covers or graphic identities, other suggestions were made. For instance, styling websites, picking adequate visual encoding and colour pallets for information visualisation, creating icons and glyphs, styling and creating combinations of clothing for fashion design, or picking musical features.

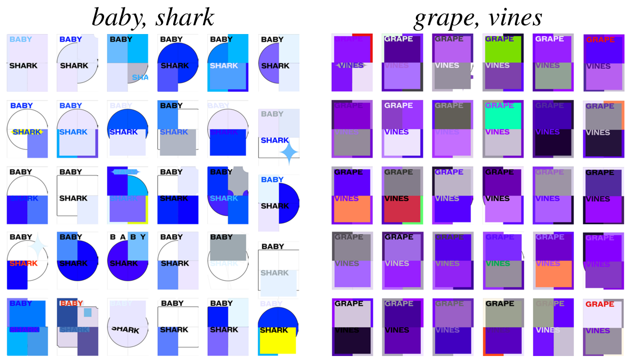


Figure 4: Posters generated from keywords that were not used for labelling mutation methods or parameters

Conclusion

More and more, designers need to be more competitive, delivering faster and cheaper, which often culminates in the adoption of trendy solutions. Thus, to speed up the exploration of innovative solutions, we presented an approach for automatically stylising graphics according to given semantic concepts, i.e. a set of keywords. To do that, *ConceptNet* (Liu and Singh 2004) was used to assess the relatedness value between given conceptual keywords and the labels of a set of mutation methods and respective parameters. The bigger the relatedness of a method, the more likely it is to be used.

We tested our approach by styling posters, i.e. transforming and styling a number of text boxes, images and geometric shapes within 2D pages. The presented approach was implemented as an extension for *Adobe InDesign* — a broadly used desktop-publishing software for GD —, so designers could alternate between manually and automatically editing/mutating the posters within the same software.

The results drawn from the user survey suggest the presented approach can be viable to aid the exploration of concept-related solutions, at least in poster design and, especially, taking into consideration that GD concepts can be transmitted in abstract ways. However, human collaboration might still be essential to curate and fine-tune the results and transform the generated ideas into final GD applications.

Future work can comprise (i) assigning weights (importance levels) to the keywords; (ii) resetting the labels according to the insights gathered through another user survey; and (iii) associating assets, e.g. if both the shape *wave* and the colour *blue* are picked because of the same keyword *sea*, then *wave* would more likely run over blue items (or neutral ones) and vice-versa.

Acknowledgments

This research was funded by national funds through the FCT—Foundation for Science and Technology, I.P., within the scope of the project CISUC—UID/CEC/00326/2020 and by the European Social Fund, through the Regional Operational Program Centro 2020 and is partially supported by *Fundação para a Ciência e Tecnologia*, under the grant SFRH/BD/143553/2019.

References

Cavanaugh, L. A.; MacInnis, D. J.; and Weiss, A. M. 2016. Perceptual dimensions differentiate emotions. *Cognition and Emotion* 30(8):1430–1445. PMID: 26308182.

Collier, G. L. 1996. Affective synesthesia: Extracting emotion space from simple perceptual stimuli. *Motivation and Emotion* 20(1):1–32.

Coyne, B., and Sproat, R. 2001. Wordseye: An automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, 487–496. New York, NY, USA: Association for Computing Machinery.

Cunha, J. M.; Lourenço, N.; Martins, P.; and Machado, P. 2020. Visual blending for concept representation: A case study on emoji generation. *New Generation Computing* 38(4):739–771.

Ha, D., and Eck, D. 2017. A neural representation of sketch drawings.

Heath, D., and Ventura, D. 2016. Creating images by learning image semantics using vector space models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Heer, J., and Stone, M. 2012. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, 1007–1016. New York, NY, USA: Association for Computing Machinery.

Lin, S.; Ritchie, D.; Fisher, M.; and Hanrahan, P. 2013. Probabilistic color-by-numbers: Suggesting pattern colorizations using factor graphs. *ACM Trans. Graph.* 32(4).

Liu, H., and Singh, P. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal* 22(4):211–226.

O'Donovan, P.; Agarwala, A.; and Hertzmann, A. 2011. Color compatibility from large datasets. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11. New York, NY, USA: Association for Computing Machinery.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents.

Rodrigues, A.; Cardoso, F.; and Machado, P. 2019. A dynamic approach for the generation of perceptual associations. In *International Conference on Computational Creativity (ICCC)*. address: organizations.

Toivonen, H., and Gross, O. 2015. Data mining and machine learning in computational creativity. *Wiley Int. Rev. Data Min. and Knowl. Disc.* 5(6):265–275.

Zhao, N.; Cao, Y.; and Lau, R. W. H. 2018. What characterizes personalities of graphic designs? *ACM Trans. Graph.* 37(4).

Prompt to GPT-3: Step-by-Step Thinking Instructions for Humor Generation

Yuetian Chen, Bowen Shi, Mei Si
Rensselaer Polytechnic Institute
110 8th street, Troy, New York 12180
{cheny63, shib5, sim}@rpi.edu

Abstract

Artificial intelligence has made significant progress in natural language processing, with models like GPT-3 demonstrating impressive capabilities. However, these models still have limitations when it comes to complex tasks that require an understanding of the user, such as mastering human comedy writing strategies. This paper explores humor generation using GPT-3 by modeling human comedy writing theory and leveraging step-by-step thinking instructions. In addition, we explore the role of cognitive distance in creating humor.

Introduction

Artificial Intelligence (AI) advancements have made significant progress in various fields, particularly natural language processing. In particular, GPT-3 has demonstrated impressive capabilities in many language generation tasks (Brown et al. 2020). This paper explores using GPT3 to generate jokes in an explainable and controlled way. As the demand for personalized and engaging content in the entertainment industry grows, there is an increasing need for AI models that can produce humorous content that resonates with human audiences. Therefore, it is crucial to develop AI models that can generate humor in a way that is compatible with human preferences.

In the field of humor generation, there are two main approaches: template-based and neural network-based methods. Template-based methods, such as those used by He et al.(He, Peng, and Liang 2019) and Castro et al.(Castro et al. 2016), rely on predefined structures that are filled with appropriate words or phrases to create jokes. While these methods are simple and easy to implement, they are limited by the predefined templates and the availability of suitable words or phrases.

On the other hand, neural network-based methods, such as those used by Zhang et al.(Zhang et al. 2020), and Akbar et al.(Akbar et al. 2021), utilize machine learning techniques. Zhang et al. used a neural network model to generate humorous captions for images by incorporating relevant knowledge from external sources. Akbar et al. fine-tuned a large pre-trained language model (GPT-2) on a joke dataset to generate jokes. These methods have the advantage of being able to learn from data and generate more diverse and original humor. However, the generation process is opaque to human users, and there is no way for people to understand how the model came up with the joke or instruct the model to generate jokes in a particular way. This may lead to the production of inappropriate or offensive humor.

This paper investigates how to enhance GPT-3's ability to generate humor in an explainable fashion by integrating Joe Toplyn – a famous late-night show writer's comedy writing theory (Toplyn 2014) through step-by-step thinking instructions to GPT3. We also explore the role cognitive distance plays in creating a humor effect.

Toplyn's Theory for Writing Jokes

Toplyn is a renowned comedy writer who has worked on shows like "Late Night with David Letterman" and "The Tonight Show with Jay Leno." He outlines his process for creating humorous content in his book "Comedy Writing for Late-Night TV: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy." (Toplyn 2014) According to Toplyn, there are four steps to crafting a joke:

1. **Create a topic sentence based on a news item:** A news article is typically used as a starting point. The first step involves generating a topic sentence highlighting the news item's main point. The aim is to create a sentence that is engaging and appropriate for humorous commentary, while still being factually accurate and not inherently funny.

2. **Identify handles and associations:** Handles refer to interesting or peculiar words or phrases found within the topic sentence, and Toplyn usually identifies two handles for each joke. After identifying handles, in the next step, associations are created. Associations are the concepts or ideas related to each handle. Toplyn generates a list of associations for each handle, which are used in the subsequent step to develop a punchline.

3. **Develop a punchline by combining associations:** A punchline connects one association from the first handle's list with another from the second handle's list. This combination should be perceived as true by most people and evoke a negative emotion towards the first major entity in the topic. The negative emotion is essential for generating humor in the monologue joke.

ISBN: 978-989-54160-5-9

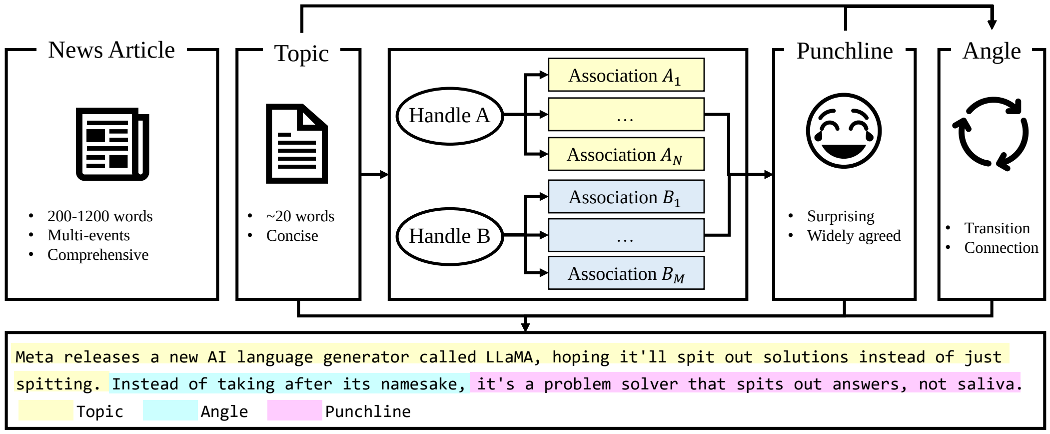


Figure 1: An overview of Joe Toplyn's theory in creating monologue jokes

4. **Connect the topic and punchline with an angle:** The angle is a sentence or phrase that smoothly transitions the audience from the topic to the punchline. It ensures that the joke has a coherent structure and flows naturally from beginning to end.

The following is an example of a joke written by Joe: "At Lax, customs inspectors seized a live shipment of 67 Giant African snails. Instead of destroying them, officials gave them jobs at the DMV." Joe Toplyn created an automated joke generation system based on his theory by prompting GPT-3 (Toplyn 2023). This work is in the same spirit. We built upon Toplyn's approach by introducing a new factor - the cognitive distance between associations for creating punchlines - to generate more effective jokes. This involves identifying seemingly unrelated associations to craft punchlines, resulting in better jokes.

Implementation and Examples

In our method, we adopt step-by-step thinking instructions that combine a chain of thought reasoning (Wei et al. 2023) with Toplyn's theory to guide GPT-3 in generating jokes. This approach includes a set of well-produced prompts for each stage of the process, allowing the model to focus on individual aspects of the joke sequentially. The chain of thought reasoning and iterative report generation during each prompt help to ensure a coherent and consistent joke creation process. The examples of intermediate results and final jokes are provided in Table 1.

Creating the Topic Sentence from News Article The instruction we used is as follows. It defines what a topic sentence is with examples:

> *"We will generate Monologue joke topics for a late-night TV show by crafting succinct sentences based on actual news items. A Monologue joke comprises three parts, and our objective is to produce the first part - the topic. The topic should be founded on a real news event that grabs people's attention and enables amusing commentary. It need not be intentionally funny but must be factually accurate. For instance, 'Carl's Jr. is selling a foot-long burger' or 'Bernie Madoff's underpants were sold at an auction' are suitable topics. During our conversation, I will provide a news article, and you will create a single sentence that fulfills these criteria. If you believe the news article is inappropriate for Monologue jokes, please inform me."*

Based on our experience, medium-sized articles (500 to 800 words) work best for generating jokes. This is because they provide enough context to write a relevant and engaging joke, while not being so long that they become overwhelming. Articles that are too short may lack enough context for a joke, while articles that are too long may contain extraneous information that dilutes the comedic focus.

Identifying Handles and Associations Handles are potential humor generators and can be people, places, things, or actions. In the prompt, we provide examples to ensure accuracy in identifying handles for a richer punchline.

> *"To further develop this monologue joke for a late-night TV show based on the provided topic sentence, we need to create the punchline, which is the surprise element at the end of the joke. To do this, follow these steps:*
>
> – *Determine two handles in the topic, which are interesting words or phrases. Handles can include people, places, things, or actions.*
> – *Brainstorm a list of associations for each handle and then create two separate lists of associations related to each handle.*
>
> *Now, based on the topic you provided, identify handles and associations in the format given by the example."*

Developing the Punchline The prompt provided below emphasizes the significance of eliciting a negative feeling towards the first major entity in the topic sentence. This is important because it creates a target for the joke, which allows the audience to laugh at the joke more easily.

> *"Pair an association from one list with an association from the other list. Choose a combination that most people would perceive as true to create the punchline. It is*

Component	Description
Topic	Microsoft introduces a new AI-powered Copilot for their 365 apps, making Clippy's ghost proud.
Handles	AI-powered Copilot, Clippy's ghost
*Associations for **AI-powered Copilot***	Artificial intelligence, **Clippy 2.0**, Microsoft 365, Productivity, GPT-4, Virtual assistant, **Automated tasks**, Office apps, Innovative technology
*Associations for **Clippy's ghost***	Nostalgia, Old technology, **Revolutionary technology**, Paperclip, **Annoying assistant**, Pop-up help, Microsoft Office, 90s tech, Failed innovation
*Punchline (**Negative**)*	**Automated tasks + Annoying assistant**: Now, it can automatically annoy you with its help.
*Punchline (**Positive**)*	**Clippy 2.0+Revolutionary technology**: Clippy's cool cousin has arrived!
*Angle (**Negative**)*	In the spirit of Clippy,
*Angle (**Positive**)*	It turning your office into a futuristic workspace, with one chatbot at a time -
*Final Joke (**Negative**)*	Microsoft introduces a new AI-powered Copilot for their 365 apps, making Clippy's ghost proud. In the spirit of Clippy, now it can automatically annoy you with its help.
*Final Joke (**Postive**)*	Microsoft introduces a new AI-powered Copilot for their 365 apps, making Clippy's ghost proud. It turning your office into a futuristic workspace, with one chatbot at a time - Clippy's cool cousin has arrived!

Table 1: Example of joke generation using the proposed method with positive and negative emotion instruction

*important to evoke a **negative** emotion towards the first major entity in the topic for the monologue joke to be humorous. Now, based on the association lists you provided, provide the punchline as shown in the example."*

Our study demonstrates that the inclusion of negative emotion in a joke plays a crucial role in its humor. Table 1 provides two instances of jokes addressing the same topic but with different sentiment prompts. When the sentiment in the prompt is changed from "Negative" to "Positive," the joke may lose its humor. This can be attributed to the fact that the joke prompted with the "Negative" sentiment keyword contains more negative emotion, resulting in greater contrast between the subjects involved.

Connecting the Topic and Punchline with an Angle In the final step, we guide GPT-3 to create a smooth link between the topic sentence and the punchline. Due to the lengthiness of the complete prompt, it is not possible to provide it here. Essentially, the previously generated content is used as input to provide context. Then, the instruction for forming the punchline is then given as follows:

"Next, craft an angle to smoothly transition the audience from the topic to the punchline."

The example included in the prompt provides guidance on how to design an effective angle. Specifically, given the topic and the punchline as follows:

```
Topic: The FBI is warning people about
the dangers of charging devices in
public areas, especially hotel lobbies.
Punchline: Your device might end up
with more than just bed bugs.
```

One of the angles could be:

```
...If you are using a USB port in a
hotel lobby, ...
```

Again, the example within the prompt reinforces the model's understanding of how to effectively weave together the various elements of the joke, transforming the topic, punchline, and associations into a cohesive and entertaining monologue joke. By combining the output generated in each stages, we can get a final report shown in Algorithm 1.

Example Outputs

In this section, we present additional examples of jokes generated using our proposed method[1].

The first example's topic is Nintendo's decision to skip E3 2023, leaving gamers curious about what they have in store. The punchline creatively links "surprise announcements" and "childhood memories," by replacing the former with the game "hide-and-seek." The resulting joke is: "Nintendo decides to skip E3 2023, leaving gamers wondering what's up their sleeve. Turns out, their new game plan is to play hide-and-seek with our childhood memories."

The second example involves Meta's new AI language generator, LLaMA. The topic centers on Meta's hope that LLaMA will generate solutions instead of just spitting. By connecting the associations of "AI language generator" and

[1]GitHub repository: https://github.com/Stry233/Joe-Toplyn-s-Monologue-Joke-generation-using-ChatGPT

Algorithm 1 Report for the new "1080p Premium" feature on YouTube

1: **Topic:** "YouTube experiments with a new "1080p Premium" option, offering higher-quality video for Premium subscribers."

2: **Associations for "1080p Premium":**

- "Premium subscribers"
- "YouTube revenue"
- "Exclusive content"
- "Higher price"
- **"Upselling"**
- "Better experience"

3: **Associations for "higher-quality video":**

- "HD resolution"
- "Better image quality"
- "More pixels"
- "Higher bitrate"
- **"Sharper image"**
- "More data"
- "Cinematic experience"

4: **Punchline:** "Upselling their way to a sharper disappointment."

5: **Angle:** "Now, viewers can pay more to witness their favorite cat videos in stunning clarity."

6: **Summary:** "YouTube experiments with a new "1080p Premium" option, offering higher-quality video for Premium subscribers. Now, viewers can pay more to witness their favorite cat videos in stunning clarity – upselling their way to a sharper disappointment."

"LLaMA," the punchline is crafted: "Meta releases a new AI language generator called LLaMA, hoping it'll spit out solutions instead of just spitting. Instead of taking after its namesake, it's a problem solver that spits out answers, not saliva."

The third example, as shown in Table 2, makes connections among three associations. It involves a Japanese spacecraft that's set to make a historic moon landing. Its cargo includes the UAE's rover and a lunar robot made by a Japanese toy maker. By combining the associations of "historic moon landing," "UAE's rover," and "Japanese toy maker," we arrived at the punchline: "As the Japanese spacecraft lands on the moon, carrying the UAE's rover and a lunar robot from a Japanese toy maker, Neil Armstrong's famous quote gets a cosmic update: 'One giant leap for UAE's rover, one small step for anime-kind'."

Create Punchlines using Unrelated Concepts

Although not mentioned in Toplyn's theory, we found that the selection of associations is a critical factor in creating engaging and humorous jokes. Selecting associations that are less obviously related can lead to more unexpected and intriguing punchlines. These types of punchlines often have a stronger comedic effect, which enhances the humor of the joke. In contrast, if we were to select associations that are more closely related, the resulting punchline may not be surprising enough to provoke laughter or amusement. In such cases, the joke might feel predictable or mundane instead.

For example, let's say we have two handles - "Space Travel" and "Fast Food." We retrieve associations for each handle as follows:

```
Space Travel:  Mars, freeze-dried meal,
astronaut
Fast Food:  burger, fries, drive-thru
```

When picking two associations for creating the punchline, we find that the pair `Mars` and `burger` are mostly irrelevant to each other. Using this pair, we can create a punchline that more likely surprises the audience and creates humor. For example:

```
Why did the astronaut bring a burger to
Mars?  Because he heard it was a great
place for a 'space'cial!
```

Alternatively, if we choose to use `freeze-dried meal` and `burger` as a word-pair which has a relatively higher semantic relevance score.

```
Why did the astronaut prefer a
freeze-dried meal over a burger?
Because it's easier to pack for space
travel!
```

This punchline lacks humor because the connection between the two concepts is easily understandable and lacks an element of surprise. The audience can quickly identify the relationship between a freeze-dried meal for astronauts and a burger, which reduces the effectiveness of the punchline in creating laughter.

Discussion and Future Work

Our work demonstrates that it is possible to use step-by-step instructions to guide GPT-3 in following human comedy writing theory, resulting in the generation of configurable and explainable jokes. The jokes are configurable because, although the process can be fully automated, humans can intervene and modify the intermediate results in the joke creation process. This advancement opens up new possibilities for AI-driven humor and entertainment and contributes to the growing body of research in the field of AI and humor studies.

The next step in our research involves incorporating additional comedy writing theories and techniques. This includes exploring various types of humor like irony, satire, and sarcasm, as well as more advanced techniques such as wordplay, puns, and misdirection. We also plan to examine the potential of integrating user feedback into the system, where users can rate and provide feedback on the generated jokes, which will be used by reinforcement learning process to improve the overall quality and effectiveness of the joke generation process.

Component	Description
Topic	A Japanese spacecraft is attempting a historic moon landing, delivering the UAE's rover and a lunar robot from a Japanese toy maker.
Handles	historic moon landing, UAE's rover, Japanese toy maker
*Associations for **historic moon landing***	**Neil Armstrong**, Moonwalk, One small step for man, Apollo, Space race, Lunar surface
*Associations for **UAE's rover***	**United Arab Emirates**, Space exploration, Desert, Arabian Nights, Sand dunes
*Associations for **Japanese toy maker***	**Anime**, Gundam, Hello Kitty, Action figures, Remote control toys, Collector's items
Punchline	Neil Armstrong's famous quote is updated: "One giant leap for UAE's rover, one small step for anime-kind.
Angle	In a cosmic twist,
Final Joke	A Japanese spacecraft is attempting a historic moon landing, delivering the UAE's rover and a lunar robot from a Japanese toy maker. In a cosmic twist, Neil Armstrong's famous quote is updated: "One giant leap for UAE's rover, one small step for anime-kind."

Table 2: Example of joke generation using the proposed method

We also plan to further study the relationship between the level of difficulty in creating connections between associations and the resulting perceived humor. Specifically, we want to investigate whether the level of cognitive effort required to make sense of seemingly unrelated associations enhances or detracts from the overall comedic effect. This could provide valuable insights into how to optimize the joke generation process for maximum comedic impact.

Finally, we want to personalize the system by considering users' humor preferences. By gathering data on users' preferred humor styles, comedians, and amusing joke types, we can train the system to create jokes that match their sense of humor. Additionally, we can try creating jokes for specific events or holidays to enhance their relevance and humor value.

References

Akbar, N. A.; Darmayanti, I.; Fati, S. M.; and Muneer, A. 2021. Deep learning of a pre-trained language model's joke classifier using gpt-2. *Journal of Hunan University Natural Sciences* 48(8).

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.

Castro, S.; Cubero, M.; Garat, D.; and Moncecchi, G. 2016. Is this a joke? detecting humor in spanish tweets. In *Advances in Artificial Intelligence-IBERAMIA 2016: 15th Ibero-American Conference on AI, San José, Costa Rica, November 23-25, 2016, Proceedings 15*, 139–150. Springer.

He, H.; Peng, N.; and Liang, P. 2019. Pun generation with surprise. *arXiv preprint arXiv:1904.06828*.

Toplyn, J. 2014. *Comedy Writing for Late-Night TV: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*.

Toplyn, J. 2023. Witscript 3: A hybrid ai system for improvising jokes in a conversation. *arXiv preprint arXiv:2301.02695*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Zhang, H.; Liu, D.; Lv, J.; and Luo, C. 2020. Let's be humorous: Knowledge enhanced humor generation. *arXiv preprint arXiv:2004.13317*.

8. Interaction and Collaboration

Recipe 2.0: Information Presentation for AI-Supported Culinary Idea Generation

Angel Hsing-Chi Hwang[1,2] **Samy Badreddine**[1] **Frederick Gifford**[1] **Tarek.Besold**[1]
[1]Sony AI
[2]Cornell University
{Angel.Hwang, Samy.Badreddine, Frederick.Gifford, Tarek.Besold}@sony.com

Abstract

We examine the effective design of information presentation (IP) in AI-supported tools to support culinary creativity. We reviewed the literature to propose a conceptual framework to guide the design of IP, such that it takes into account key constructs and processes of idea generation (i.e., the motivation, subject, and content of information presented to facilitate ideation). In Part 1, we collected feedback from professional chefs regarding concerns and needs around AI tools for culinary innovation. In Part 2, we performed a content analysis on popular culinary content that inspired users to ideate. In Part 3, we designed interactive prototypes based on these insights and conducted a large-scale user study (N = 250). We found that showing the cause-and-effect logic of cooking by demonstrating information in a "what if..." format encouraged new ideas. Novice users were most motivated by understanding the ingredient' constraints and learning cooking practices' rationales. Through this work, we present design implications for AI-supported idea generation and call for more efforts to study how AI can augment human creativity and other open-ended decision-making processes.

Introduction

As artificial intelligence (AI)—e.g., in the form of generative systems—increasingly enters the creative space, we see the need to establish more effective and sustainable forms of human-AI collaboration (Gillies et al. 2016; Kamar 2016). To contribute to this vision, we see ***AI-supported idea generation*** (ASIG) as a promising area for further exploration:

1. Though the majority of work in AI-supported decision-making focuses on building tools that attain more accurate and optimal solutions, abundant decision scenarios are open-ended in nature and do not necessarily entail a single best solution. Understanding how AI can support idea generation can contribute to the latter type of decision-making, which remains under explored.
2. Building on (1.), we see idea generation as an avenue where users can bring their unique strengths in the decision process, while AI offers support that may be hard to attain by humans themselves (e.g., synthesizing and presenting large amount of data). In such dynamics, we see genuine collaboration between the two parties.
3. Compared to obtaining creative end products directly from AI, supporting users at the earlier stage of a work process (i.e., during idea incubation) allows more input and control from users, encouraging a more engaged form of human-AI interaction.

We study ASIG for culinary innovation which allows us to examine the research topic through a number of critical nuances. Both users and AI can bring in unique contributions during the work process, forming authentic collaboration. For instance, AI can summarize and present large amounts of food data (e.g., recipes, nutrition or chemical information of ingredients, as well as food waste and the environmental impact of cooking) which may be difficult for users to collect and digest on their own. Meanwhile, human cooks hold a number of decision factors that may not otherwise be shared by the machine, such as experience of tastes, context for cooking, cultural background, and personal preferences. Taking these human-centered factors into account can inform how AI can support resolving open-ended questions where humans' subjective values – instead of standard metrics – are used for evaluation. One can produce a dish that scores high on nutritional benefit and low on environmental impact, but the success of the cooking idea ultimately depends on whether it tastes good to human diners.

As an application domain, compared to other more specialized practices (e.g., healthcare, legal, finance), cooking covers a wide spectrum of decision-makers, ranging from highly skilled chefs to casual home cooks. Upon reaching an idea, users themselves have to execute the cooking process, meaning this topic requires them to be more involved in the decision process irrespective of expertise.

As we reviewed prior work and theoretical foundations, we also see cooking as a suitable domain to address the various components and cognitive processes of idea generation. Against that backdrop, we conducted a three-part research process to examine the use of AI for culinary idea generation. First, we conducted a formative interview with professional chefs to grasp their opinions toward ASIG tools for culinary innovation. Next, to also seek insights from more diverse communities and general users, we took a computational social science approach to analyze key features in effective online culinary content (e.g., cooking tutorials) that inspire cooking ideas. Based on insights from our content analysis, we designed five different prototypes of interac-

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

tive tools that support cooking idea generation, and we conducted a large-scale user study to collect user feedback and design implications.

Our work makes practical and theoretical contributions along three dimensions. First, based on a literature review, we synthesize a framework with key constructs of idea generation and show how it can be adopted to inform design of ASIG tools, leading to more fruitful ideation outcomes than those of existing tools. Second, we assemble perspectives from diverse communities, providing concerns that need to be addressed and directions for improving emerging technologies in the culinary domain. Finally, we examine outcomes of taking these directions into actual design, as well as their effect on users' experience levels. We identify the importance of informing users about the logic and constraints in idea generation, and show that these design practices are particularly helpful to inspire new ideas among novice users. In sum, our work contributes to exploring AI's potentials in culinary innovation, while expanding the broader knowledge space of AI's applications to more open-ended, personalized, and creative decision-making.

Background and Related Work

Cognitive Constructs of Idea Generation

While idea generation in each domain applies unique expertise and practices, existing research identifies several universal principles and elements that are crucial to the process of generating and formulating ideas (Girotra, Terwiesch, and Ulrich 2010; Toubia 2006). Most previous work builds on the notion that creative ideas need to be both "new and appropriate" (Sternberg and Lubart 1999; Kaufman and Sternberg 2010). For idea generation, appropriateness needs to take into account the capabilities and constraints of the involved materials and methods (Medeiros et al. 2018; Rietzschel, Slijkhuis, and Van Yperen 2014). Knowing *"what works"* and *"what doesn't work"* become equally important.

Then, how does one progress from ideas that are simply appropriate to those that are truly innovative? Design researchers and practitioners have proposed outlining the governing logic and then encouraging individuals to adapt and apply it to their own ideation problems (Plucker 2004; Policastro and Gardner 1999). In other words, knowing *"how things work"* enables one to apply functional practices across domains and topics, leading to ideas that are out-of-the-box but still possible to execute. Additionally, research in cognitive science found inhibitory control to be a crucial cognitive function for idea generation (Cassotti et al. 2016; Flaherty 2005). Specifically, humans have the tendency to adopt cognitive heuristics (i.e., mental "shortcuts" that allow one to make decisions and take actions quickly). This indicates that ideas that are more common and obvious often come to mind first during idea generation. Therefore, whether a person can suppress these highly accessible, "easy" ideas becomes decisive for innovation.

Putting these considerations together, a handful of research looked at the process of idea generation, which typically entails the following stages: (1) identifying opportunities and problems, (2) acquiring knowledge and collecting information, (3) generating preliminary ideas, (4) evaluating and further developing ideas, and (5) implementing, revising, and improving ideas (Shneiderman 2002; Treffinger, Isaksen, and Stead-Dorval 2006). Throughout these stages, individuals are said to alternate between two key cognitive approaches: *divergent thinking* and *convergent thinking*. With divergent thinking, one would cast a wide net, trying to collect as much information, identify as many opportunities and gaps, and lay out as many potential ideas as possible (Runco and Acar 2012; Acar and Runco 2019). Divergent thinking is often seen as an indicator of creativity – after all, starting broad would provide a greater number of "candidate" ideas for a person to further develop, which, again, can lead to higher quality of ideation outcomes. However, not all working materials and initial ideas are worth further pursuing and developing, and some may not even be feasible for execution. Convergent thinking narrows down the scope, prioritizes what may be more relevant, and identifies novel problem space (Cropley 2006; Simonton 2015). In particular, convergent thinking supports evaluation, a critical step that allows one to focus on more promising ideas and further develop them, leading to true innovation. Therefore, the use of both divergent and convergent thinking is common in popular ideation techniques, such as brainstorming (Larey 1994), Linkographs (Goldschmidt 2016), and the Double Diamond model (West et al. 2018).

AI-Supported Idea Generation

Existing work in computational creativity has demonstrated AI's potential in automatically and independently generating creative content as well as in augmenting human creativity through the provision of tools or creative collaborators (Davis et al. 2015). Here, we focus on the latter and discuss several cohesive themes from proposed approaches to designing ASIG tools.

AI tools are supposed to ***align well with users' mental and work processes of idea generation***; for instance, (Schleith et al. 2022) proposed to use six key actions (i.e., learn, look up, relate, monitor, extract, and create) as creative prompts to guide users through the ideation process and land on novel ideas. Also, in order to elicit ideas above the ordinary, various studies emphasize the importance of ***creating interactive experiences***. Drawing from work in social robotics, one common approach is to "bounce ideas back and forth" with AI tools, enabling users to take turns and shift initiative between themselves and the tools (Lin et al. 2020; Gero, Liu, and Chilton 2022). Such experiences—especially when AI tools provide unexpected content—were found to spark inspiration and unblock ideation bottlenecks.

Another successful strategy is to ***provide users with more opportunities to collect feedback (and thus evaluate their ideas)***, especially when giving users control to customize the types of feedback to their needs. For instance, (Wu, Terry, and Cai 2022; Wu et al. 2022) created a writing support tool and a music composition tool with large language models while allowing users to create their own interactive experience through prompt chaining. This allowed users not only to better understand how the AI tools worked but also how they could improve their ideation content through more personalized, granular feedback.

Finally, several studies stress AI's capabilities of ***integrating and presenting large amounts of information*** to help users acquire knowledge and enrich sources of inspiration.

Information Presentation (IP) to Support Ideation

We see great promise in the capacity of AI systems to extract information from (often large amounts of) data. To harness this strength, we focus on how AI tools can support idea generation through effective information presentation (IP). How individuals encode information can impact whether they can leverage it for idea generation and creative problem-solving (Mumford et al. 1996; Sawyer 2011). Specifically, presentation that can help users focus on factual information, discount the irrelevant, and connect the dots, can drive higher-quality ideas (Mumford et al. 1996; Mobley, Doares, and Mumford 1992). (Wang and Nickerson 2017) studied tools and systems that support creative work through assembling and presenting information from digital libraries and the web, focusing on task-specific knowledge, and enabling more efficient information search. The review found effective creativity support systems often serve at least one of three functions: structure and organize knowledge hierarchically, synthesize and provide various perspectives to an existing topic, and filter and offer information based on its relevance.

Various approaches to implementing these functions have been examined for their effectiveness on creative problem-solving and ideation. Early work attempted to present information step-by-step depending on users' different stages of design thinking or work processes (Elam and Mead 1990; Marakas and Elam 1997). Alternatively, (Althuizen and Wierenga 2014; Forgionne and Newman 2007) focused on offering concrete case studies and examples, in the hope that users could draw analogies between these references and their own work as inspiration for new ideas. (Wang and Ohsawa 2013; Jenkin et al. 2013) designed tools to extract and offer key notes from large amounts of information, directly highlighting important and novel points for users. Across the board, visualization was found as a particularly helpful means of information consumption (Kohn, Paulus, and Korde 2011).

Beyond idea generation, IP is critical to designing AI tools for decision support in other domains. Besides addressing common challenges about users' trust in AI and its explainability (Goebel et al. 2018), recent work has revealed a lack of actionability as a key drawback of such systems (Yang, Steinfeld, and Zimmerman 2019). To address this issue, we construct a framework to guide designers of ASIG tools to present information while taking into account the key cognitive constructs and processes of idea generation discussed above. Our proposal entails the following three layers:

- Motivation (the "why" problem): Whether the goal of IP is to ***support divergent or convergent thinking***; namely, whether the tool should help users explore a wide range of relevant knowledge, references, and examples (divergent thinking), or focus on just one or a few sample(s) to help users funnel their thoughts to a specific end.
- Subject (the "what" problem): To inform the capabilities and constraints of ***materials and methods*** that one works with for idea generation, ASIG can present information about materials, methods, or a combination of both.
- Content (the "how" problem): To put information into tangible IP, one should address one or several of the following key questions: ***"what works", "what doesn't work", and "how things work"***. For instance, one can provide details for a material or a method that users work on (i.e., ***input of ideation***); one can explain the logic of how things work, inspiring users to generalize and apply them in another domain (i.e., ***rules of ideation***); or one can show possible outcomes and examples as sources of inspiration ***output of ideation***.

Technology for Culinary Innovation

Prior human-computer interaction work has explored various approaches to augment culinary innovation, including enhancing social engagement in cooking experiences (Isaku and Iba 2015; Svensson, Höök, and Cöster 2005) or generating 3D artifacts of food to serve as sources of inspiration (Sauvé, Bakker, and Houben 2020; Naritomi and Yanai 2021; Punpongsanon et al. 2022). Other work looked at supporting users to collect helpful information (e.g., nutrient data, cooking techniques, cookware, and recipes) in order to come up with novel cooking ideas (Baurley et al. 2020; Yoneda and Nadamoto 2018; Kato and Hasegawa 2013). Here, one of the key challenges lies in the divergent and highly subjective notions of what users consider as useful and meaningful data (depending, e.g., on taste preferences, dining experiences, and cultural background). Correspondingly, efforts have been made to build more personalized recommendation systems (Chen et al. 2021) as well as IP and visualization tools that effectively reveal insights and spark inspiration for cooks (Chang et al. 2018). Another approach to leveraging insights from food-related data is through directly generating new cooking ideas, e.g., by providing food pairing recommendations (Gim et al. 2022), by suggesting how food ingredients and cooking methods pair well together (Baurley et al. 2020; Yoneda and Nadamoto 2018; Kato and Hasegawa 2013), or by generating entire new recipes from food tutorial clips (Fujii et al. 2020). Finally, across the various approaches to supporting culinary innovation, there is a growing trend to adopt and present information from multiple domains, e.g., simultaneously providing information about food and its environmental impact (Kuznetsov, Rodriguez Vega, and Long 2022; Sauvé, Bakker, and Houben 2020).

Several of these IP approaches have been implemented and productized; for instance, recipe recommendations and food pairing functions are shown in various commercial applications (e.g., BigOven, PlantJammer, FoodPairing). However, these applications face the challenge to fit well and embed into users' existing cooking practices. In particular, it is often required that users possess clear cooking goals prior to starting information collection. This counteracts with more common approaches to creative ideation, which more often start from broader, divergent scopes, and later converge to more specific paths (Urban Davis et al. 2021). Moreover, while existing literature has explored a variety of materials and formats targeting specific cooking elements (e.g., ingredient, cookware, time, temperature) to support culinary innovation, more principled guidelines to inform this design space are lacking.

Regarding the general study of ASIG, creativity-support tools for cooking offer an interesting application domain in that the previously described theoretical constructs of idea generation can be practically operated and assessed:

- Motivation (the "why" problem): Divergent approaches to present culinary information would demonstrate the various dishes that an ingredient or technique can be applied to, while convergent reasoning would highlight how information can be applied to a specific dish.
- Subject (the "what" problem): Under the context of cooking, subjects that users work on to generate new ideas include ingredients (materials) and cooking techniques (methods).
- Content (the "how" problem): For cooking, one can focus on presenting general good practices to cooking (i.e., rules), informing specifics of an ingredient or technique (i.e., input), or showing examples of cooking outcomes (i.e., output).

In the following we examine how different approaches to IP can support generation of creative cooking ideas. We begin with understanding users' perspectives on existing AI tools for culinary ideation to identify strengths and pain points of these tools and possible means for improvement. While the majority of these creativity-support tools target culinary professionals, we further ask whether their current advantages and disadvantages are applicable to general users and how adaptation can be achieved. Taking these general directions for improvement and/or adaptation, we then study how to execute them into actual content and design. Here we take a computational social science approach to observe and pursue inspirations on the public discourse. Specifically, we perform content analysis on popular cooking tutorials and their audiences' responses in order to find content and design strategies that generally work well and elicit creative cooking practices. Finally, informed by the content analysis, we execute effective content design into five interactive prototypes and launch a large-scale user study to examine users' behavior, experience, and feedback.

Part 1: Perspectives on Existing Tools

A wide range of technologies for culinary innovation have been designed and targeted at expert users (i.e., professional chefs). As well as a general interview series with eighteen chefs, and in-depth interviews with six chefs and two industry advisors, we conducted a semi-structured, formative interview with 16 professional chefs from Europe, North America, South America, and Asia to capture their feedback on the advantages and drawbacks of some existing tools, specifically comparing effective recommendation and food pairing functions. Here, we summarize insights that are relevant to the current work.

First, experts favored the specificity of information; for them, it is the details that matter and serve as effective inspiration. For instance, when providing information on an ingredient, it is important to specify its origin and processing techniques. Also, presenting information about *either* ingredient *or* method was seen as an ineffective approach to professional cooks. In general, chefs favoured a convergent approach to IP; they showed little interest in seeing a large amount of general information or its summary, as most of it usually seemed already familiar and less inspiring to them. Instead, they preferred focusing on specific, commonly unseen ingredients or techniques as sources of inspiration. Experts' opinions on IP can be summarized as follows:

- Motivation (the "why" problem): Adopt a ***convergent approach*** and present mainly selective, highly relevant, and previously unseen information.
- Subject (the "what" problem): Present information about both ***materials and methods*** side by side.
- Content (the "how" problem): Present detailed information about the ***input*** for idea generation.

Are these insights applicable to designing ASIG tools for general users? Generally, layman users often lack specific goals when initiating idea generation processes (Sawyer 2011). Taking a convergent rather than divergent approach to design ASIG tools may therefore not be as effective for non-experts. Furthermore, jointly presenting different types of information (i.e., combining ingredient- and method-related information) may cause information overload for non-experts, as they are less apt at connecting the dots across abundant information than their expert counterparts (Mumford et al. 1996; Mobley, Doares, and Mumford 1992). Finally, providing highly specific, detailed information to general users may also reduce the flexibility for innovation (Kletke et al. 2001). Unlike professional chefs, who know how to substitute one ingredient with another, amateur cooks may not know the alternatives to specialized items. Therefore, in the following, we gathered insights from a broader, more general user base.

Part 2: Effective Content to Support Ideation

Method

To get insights from general users and to see how IP can be executed in effective content design to inspire ideation, we took a computational social science approach to review popular culinary content on YouTube and corresponding audience responses. We chose this medium for content analysis as cooking tutorials on YouTube are one of the most common resources where the general population seek cooking information (Benkhelifa and Laallam 2018). We focus on understanding what types of content—as well as their approach to IP—can elicit more creative ideas with lay users.

Cooking videos for content analysis We examined the top 10 most popular channels under YouTube's food and cooking category, and used `BeautifulSoup` from Python to crawl the links and metrics of the top 10 most viewed cooking videos in each channel and the top 100 most engaged comments for each video. In total, we obtained 100 cooking videos and 10,000 comments for the analysis.

Coding approach To analyze the video content and understand users responses to the culinary information presented in these videos, we first viewed the videos and

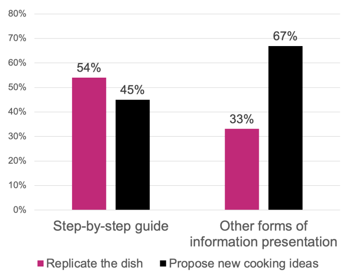


Figure 1: Viewers' responses to popular cooking videos

remarked how information was presented in each video through our proposed framework. We then examined the audience feedback to each video through two specific lenses: whether viewers mention cooking ideas and/or actions taken upon viewing the video, which helps us investigate if the video was effective in triggering idea generation; and how viewers described the content and what they found beneficial from each video, which allows us to distill features for effective content design for culinary innovation.

Results

Among the examined culinary videos, a considerable portion (28%) simply followed a conventional, step-by-step tutorial style to show viewers how to cook a dish in a specific way from the start to the end. The remaining videos demonstrated cooking information through various formats, such as doing experiments to show how switching ingredients and applying different cooking techniques may influence cooking outcomes. The majority of these cooking tutorials applied convergent approaches (76%), featuring the production of only one dish in each tutorial, and covered both input (specified information about ingredients: 69% of all videos; specified information about cooking techniques: 88% of all videos) and output of cooking (showed results of cooking: 70%). Furthermore, as many as 72% of videos applied examples to present culinary knowledge. In the contrary, IP emphasized "what worked" for each dish (79% of all videos), while the logic of "how things worked" (9%) and "what didn't work" (32%) were less mentioned.

Examining feedback to the cooking videos, we noted different patterns in viewers' comments depending on how culinary information was presented. Whether or not a video used step-by-step cooking instructions (i.e., presenting one way to "do the dish right") has a salient effect on viewers' ideas and intended actions. After watching a video showing how a dish was made from the start to the end, viewers were more likely to express interest in making the dish (mentioned among 41% of comments) with more than half (54%) intending to replicate what they learned from the tutorial step-by-step. Two-thirds (67%) of those who watched videos presenting culinary information in more diverse ways would instead propose their new ideas for cooking.

We examined IP formats that inspired viewers to generate their own cooking ideas. Viewers responded particularly positively to three types of content. First, the audience was interested in ***understanding the "cause-and-effect" rationales*** behind cooking. Participants proposed more new cooking ideas after watching videos that asked numerous "what if..." questions and demonstrated how changing one component in cooking (e.g., varying a cooking technique, ingredient, or cookware) would influence the outcome of the dish (e.g., flavor profile, texture). For instance, in one of the videos, the chef varied the time and temperature to sear a piece of steak and examined its tenderness. Experimenting and showing how different methods led to distinct outcomes was seen as particularly informative and motivated viewers to come up with their own plans to create steak dishes that suited their taste preferences. Secondly, viewers prefer information that ***advises how they can adapt what they learned*** from a cooking video to their own kitchen. Understanding how a cooking technique can be generalized to handle different ingredients and produce various dishes was, thus, found especially helpful and triggered more proposals of new cooking ideas. Third, ***presenting constraints*** of an ingredient or method is valuable. As mentioned in viewers' comments, this information helped them to understand what may have gone wrong in past cooking attempts and to come up with new ideas avoiding those mistakes. Likewise, seeing chefs' trial-and-error processes was perceived as helpful to comprehend what would not work and to come up with ideas that could make a cooking plan successful. Moreover, these three approaches to presenting culinary information not only encouraged idea generation, but they were also reported to enhance the positive effect—such as fun and enjoyment—of cooking processes, which is yet another motivating factor of innovation (Sawyer 2011).

Summarizing our findings from studying general audience comments on publicly available cooking videos:

- Motivation (the "why" problem): Public comments showed no particular preference between presenting information in either convergent or divergent approaches.
- Subject (the "what" problem): Public comments showed no particular preference between understanding information about materials and methods.
- Content (the "how" problem): Viewers expressed particular interest in understanding the cause-and-effect (i.e., explaining rules and "how things work"), adaptability (i.e., explaining rules, "how things work", and "what works"), and common mistakes in cooking (i.e., information related to constraints of input and "what doesn't work").

Part 3: Designing AI Tools to Support Culinary Ideation

Method

Based on insights from the previous sections, we designed five different prototypes, each applying a unique IP strategy to support culinary innovation. In Table 1, we summarize how each design of the five conditions corresponds to key constructs of idea generation. We then conducted a between-subject user study to examine whether and how users' performance differs when adopting different idea generation tools. Participants read an informed-consent form and completed a pre-survey on their cooking practices and habits. As a key part of the study asks participants to gen-

Table 1: Approaches to information presentation (IP) and their corresponding outcomes

Insights (Part 1 & 2) or Condition (Part 3)	Motivation of IP (The "Why" Problem)	Subject of IP (The "What" Problem)	Content of IP (The "How" Problem)	Show "what works"	Show "what doesn't work"	Explain "how things work"	Show examples	Ideation Outcomes	Affected Users
Feedback from experts (Part 1)	Support **convergent thinking**	Material + Method	Input	✓				(Not applicable)	(Not applicable)
Popular content on public discourse (Part 2)	Support convergent thinking: 76% Support divergent thinking: 24%	Material: 69% Method: 88%	Input: 95% Output: 70% Rule: 32%	79%	9%	29%	72%	(Not applicable)	(Not applicable)
Feedback from audiences' comments (Part 2)	(Not available)	No noticeable preference	Input + Rule (the cause-and-effect, adaptability, and common mistakes)	✓	✓	✓		(Not applicable)	(Not applicable)
"Baseline" condition (Part 3)	Support **convergent thinking**	Material + Method	Input	✓				(+) for revising ideas (-) for generating new ideas	Encourage experts to generate new ideas but lead novice users to copy
"Pairing" condition (Part 3)	Support **divergent thinking**	Material	Input	✓				(+) for revising ideas (-) for generating new ideas	Encourage novice users to revise existing ideas
"Generalizable Method" condition (Part 3)	Support **divergent thinking**	Method	Rules + Output	✓		✓	✓	Encourage new ideas but also replication	No particular affected users
"Constraint" condition (Part 3)	Support **divergent thinking**	Material	Rules		✓	✓		Encourage new ideas but also replication	Encourage novice users to generate new ideas and revise existing ones
"What if" condition (Part 3)	Support **divergent thinking**	Material + Method	Rules + Output	✓		✓	✓	Enhance generation of new ideas	Encourage most users, especially those with more experiences, to come up with new ideas

erate a cooking idea, we used the pre-survey to screen and exclude participants who had no cooking experience.

Participants were randomly assigned to one of the five prototypes, to explore, use and come up with cooking ideas. While exploring, they composed a recipe plan, which entails describing a dish they would like to create, the ingredients needed, and the steps they would take to cook the dish. Participants were asked to pull up and view their assigned prototype alongside their recipe planning screen. They did not have to memorize and could instead refer to information in the prototype when generating ideas. As last step, participants filled out a short exit survey to reflect their user experience and reported their demographic data. The entire study took around 30 minutes to complete, and participants received $10 to compensate their time and participation. The study was reviewed and approved by the Research Ethics Board at the authors' affiliation.

Interactive prototypes We used Figma to create interactive prototypes for the five conditions of our user study. The designs of the conditions were informed by the findings we reported in Part 1. To create equivalent initial states, all conditions started with showing users a home page with four dishes as sources of inspiration. Participants could click to explore each dish. From there, each prototype applied a unique strategy to present information and inspire idea generation. The five conditions include: (1) **baseline condition**: a classic step-by-step recipe; (2) **ingredient pairing**: a recipe showing the molecular and recipe fit of ingredients used; (3) **constraint condition**: a recipe showing the constraint of each ingredient used; (4) **generalizable method condition**: a recipe showing the cooking techniques used and other cuisines that can be made applying the same methods; (5) **what-if condition**: a recipe showing possible outcomes as one switches the ingredients and cooking techniques. Each condition had a similar amount of information for exploration with each dish having 6 pages of content to click through and each condition having in total 24 pages of information to consume. To check if participants had explored the content on their assigned prototypes, we included a page code on each piece of the content, and participants were asked to record and report the code of pages they had viewed in the exit survey. None of our participants failed this validity check.

Measures of the user study We collected three main categories of data, including users' existing cooking practices and experience (frequency, expertise, and years of cooking), the recipe idea they planned out, and their user experience during idea generation (measured through the usability scale (Bangor, Kortum, and Miller 2008) and the self-efficacy scale (Sherer et al. 1982)). With participants' cooking plans, we coded each idea into one of three types: (1) *copy* indicates a participant was simply copying the idea from one of the four source dishes; (2) *revision* indicates participants adopted one of the four source dishes but made a twist to its original recipe (e.g., swap ingredients or replace a cooking technique); (3) *new* indicates a participant came up with their own cooking idea that are distinct from the four sample dishes presented in the prototype.

Participants We recruited 250 participants on Amazon Mechanical Turk (AMT) through the following screening criteria: participants were located in the United States and have completed more than 1000 HITs with a HIT approval rate greater than 98% at the time when the study was conducted. Average age of participants was 36.30 ($S.D. = 11.17$). The majority of participants was Caucasian (70%), while 54% identified as female and 40% as male. As stated in our recruitment message and research consent form, participants also needed to have at least one month of cooking experience at the time they participated in the study. Overall, participants have, on average, 16.29 years of cooking experience ($S.D. = 11.63$). The final sample size was predetermined by conducting a pilot study and performing a power analysis based on the pilot data.

Results

We saw that **the types of recipe ideas participants came up with differed significantly by the prototype conditions** ($\chi^2 = 23.898$, $p = 0.002$). Specifically, those who explored the *what if* condition generated the most new recipes (76.32%), followed by those who explored *generalizable methods* (55.81%) and *constraints* (56.10%) of ingredients. Participants who viewed the *baseline* (47.73%) and the *ingredient pairing* conditions (40.48%) came up with the fewest new ideas. Still, although not inspiring idea generation, the last two prototypes seemed effective in providing information that can be useful for adaptation. Correspondingly, we saw the largest portion of participants who revised recipes from the four source dishes (baseline: 25.00%; ingredient pairing: 26.19%). Also in their descriptions of recipe plans we saw the highest percentages of mentions of how participants were able to apply learned information for the *baseline* and *pairing* prototypes.

We also ascertained the degree of complexity of participants' recipe ideas by examining the number of ingredients and the number of planned out steps. The number of ingredients used in different conditions differed marginally by conditions ($F = 2.043$, $p = 0.089$). Participants who viewed the *baseline* condition ($M = 9.43$, $S.D. = 4.61$) applied the most ingredients, followed by those in the *what if* condition ($M = 9.03$, $S.D. = 3.80$), and those in the *ingredient pairing* condition ($M = 8.20$, $S.D. = 3.80$). There was no significant difference either in the number of steps planned out in participants' cooking ideas ($F = 0.502$, $p = 0.735$) or in participants' self-reported user experience and ease of use in the different prototypes explored ($F = 1.672$, $p = 0.158$). This rules out the alternative explanation that participants were more likely to come up with their own cooking ideas simply because they couldn't acquire or comprehend the sources of inspiration in a prototype.

We used the number of years participants spent cooking multiplied by their cooking frequency as a proxy to assess their cooking experience levels. Overall, we saw a marginal effect of participants' cooking experience on their idea generation outcomes ($\beta = 0.02$, $S.E. = 0.01$, $t = 1.68$, $p = 0.095$). Moreover, **participants' existing cooking experience moderates their ideation outcomes**, as we found an interaction effect between participants' experience levels and the recipe condition they explored ($\beta = -0.03$, $S.E. = 0.02$, $t = -1.79$, $p = 0.075$). To be specific, for the *what if* or the *baseline* prototypes, more experienced participants came up with more new ideas; conversely, novice participants were triggered to generate more new ideas when they explored the other conditions.

Discussion

We reviewed prior literature and synthesized a theoretical framework to guide the design of IP in ASIG tools. Specifically, we proposed that effective IP should cover three key constructs of idea generation, responding to the "why" problem (i.e., whether the motivation is to support divergent or convergent thinking), the "what" problem (i.e., whether the subject entails the material or method of idea generation), and the "how" problem (i.e., whether the content addresses "what works," "what doesn't work", and "how things work" through providing information about the rules, input, or output of idea generation). We first obtained perspectives from experts regarding the benefits and shortcomings of existing tools, sought insights from the public discourse on effective content that can inspire ideation, and conducted a user study to examine the effectiveness of different interactive prototypes. We now elaborate on the various theoretical and practical implications we gather from the results.

Design Implications

The different patterns in viewers' responses to differing styles in popular culinary content reconfirmed the important roles of IP in inspiring idea generation and creative problem-solving (Mumford et al. 1996; Mobley, Doares, and Mumford 1992). Specifically, upon viewing more of a constrained, conventional style, presenting step-by-step guides to "do things right in one specific way", the audience were more likely to replicate the same course of actions instead of coming up with their own cooking approaches and ideas. As regards the design of IP that can effectively inspire ideation, we found that all three of the newly proposed interactive prototypes (*generalizable method* condition, *constraint* condition, and *what if* condition) led to more new ideas generated than what was the case for approaches seen in existing tools (i.e., *baseline* and *pairing* conditions).

In view of similar results in the *baseline* and *pairing* conditions, we realized that the shift from adopting convergent to divergent approaches per se did not necessarily lead to different idea generation outcomes. What seemed to matter more is what information was covered and how IP was designed in the actual content. Regarding IP, we first compare the outcomes of the *generalizable method* condition and *what if* condition, as these two conditions differed only in the subjects presented, with all other design factors following similar strategies. Having more new ideas generated in the *what if* condition suggests covering both the material and method as subjects of IP has a additive effect on idea generation. However, while covering two subjects at once may be informative for more experienced users, we saw that those with fewer cooking experiences in these cases ended up copying more. We also saw a similar pattern in the *baseline* condition, which also jointly covers ingredients and cooking techniques. Indeed, as seen in Part 1, professional chefs particularly requested to see detailed information about ingredients and techniques side by side. Still for lay users this design may introduce too much information at once therefore becoming less helpful.

The general public especially expressed interest in understanding the logic behind "how things work" and identifying common mistakes. We found that less experienced users performed particularly well in generating new ideas when they worked with the *constraint* condition. This highlights the effectiveness of explaining applicable rules, reasoning, and limitations of an ideation subject in IP, while, according to the data from Part 2, these design strategies were less often applied in existing content. Judging from our literature review, they are also less emphasized in prior work on

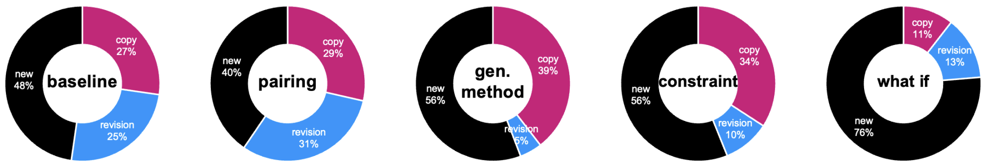


Figure 2: Breakdown of recipe ideas generated in each prototype condition

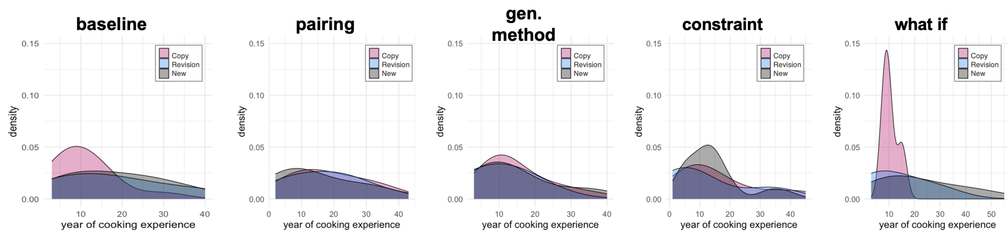


Figure 3: Density distribution of idea generation outcomes by participants' years of cooking experience in each condition

designing ASIG tools. It is also worth noting that—while the literature often emphasizes the usefulness of presenting examples as sources of inspiration—our findings from the *constraint* condition suggest that the absence of examples at least did not hinder users from coming up with new ideas.

Limitations and Future Work

We acknowledge several limitations of our current work. First, to compare how different designs of IP affects users with higher versus lower levels of experience, we used participants' years of cooking experience as a proxy to their experience level. This may not be rigorous enough and can be subject to other confounding factors. For instance, a user may have had formal culinary training and be more skilled than a home cook who has spent many more years cooking. Ideally, we would want to bring the interactive prototypes to professional chefs and conduct user studies with them as well. We encourage future work to perform more rigorous comparison between experts' and novices' responses.

Second, while we examine whether participants came up with new ideas as a key measure, we acknowledge that the nature of idea generation is much more complex and should be further explored through multiple dimensions. For instance, although participants in the *baseline* and *pairing* conditions did not generate as many new ideas, they tended to work well on revising ideas. This may respond to the general public's interest in adapting what they learned to their own cooking environment. At the same time it also suggests these forms of IP may facilitate users to learn and absorb knowledge. Because a positive relationship between learning and creativity has been found in the long run, we might observe different effects of the two conditions if we extended the study period over a longer time span. This is another potential direction which we encourage future research to pursue.

We asked each participant to plan out just one recipe idea. We adopted this approach to focus on investigating the quality instead of the quantity of participants' idea generation outcomes. Still, participants thus did not have the opportunity to compare, evaluate, and select the best idea out of a pool of candidate ideas they generated. While evaluation serves as a critical component in the full process of idea generation, we see the need for additional work to examine effective IP to support users' evaluation and selection of ideas.

More broadly our current work contributes to understanding how humans work with AI to resolve open-ended problems that rely on users' subjective, personal experiences in decision-making processes. Building on our findings, an important next step will be to compare how effective design of IP to support open-ended decision-making differs from information presented for close-ended decision-making.

In summary, our present work is an initial attempt to propose more systematic approaches to designing IP in ASIG tools. We use culinary innovation as a domain to operationalize our theoretical framework and conduct user studies, while we encourage future work to examine the topic across (and adopt our proposed framework to) other specialty areas. Additionally, collecting information is just one of the various steps in idea generation; designing AI tools to support other parts of the ideation process still remains an underexplored area offering itself for further research.

References

Acar, S., and Runco, M. A. 2019. Divergent thinking: New methods, recent research, and ext. theory. *Psychology of Aesthetics, Creativity, and the Arts* 13(2):153.

Althuizen, N., and Wierenga, B. 2014. Supporting creative problem solving with a case-based reasoning system. *Journal of Management Information Systems* 31(1):309–340.

Bangor, A.; Kortum, P. T.; and Miller, J. T. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24(6):574–594.

Baurley, S.; Petreca, B.; Selinas, P.; Selby, M.; and

Flintham, M. 2020. Modalities of expression: Capturing embodied knowledge in cooking. In *Proc. of the 14th Int. Conf. on Tangible, Embedded, and Embodied Interaction*, TEI '20, 785–797. New York, NY, USA: ACM.

Benkhelifa, R., and Laallam, F. Z. 2018. Opinion extraction and classification of real-time youtube cooking recipes comments. In Hassanien, A. E.; Tolba, M. F.; Elhoseny, M.; and Mostafa, M., eds., *The Int. Conf. on Advanced Machine Learning Technologies and Applications (AMLTA2018)*, volume 723, 395–404. Cham: Springer International Publishing.

Cassotti, M.; Agogué, M.; Camarda, A.; Houdé, O.; and Borst, G. 2016. Inhibitory control as a core process of creative problem solving and idea generation from childhood to adulthood. *New directions for child and adolescent development* 2016(151):61–72.

Chang, M.; Guillain, L. V.; Jung, H.; Hare, V. M.; Kim, J.; and Agrawala, M. 2018. Recipescape: An interactive tool for analyzing cooking instructions at scale. In *Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems*, CHI '18, 1–12. New York, NY, USA: ACM.

Chen, Y.; Subburathinam, A.; Chen, C.-H.; and Zaki, M. J. 2021. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In *Proc. of the 14th ACM Int. Conf. on Web Search and Data Mining*, WSDM '21, 544–552. New York, NY, USA: ACM.

Cropley, A. 2006. In praise of convergent thinking. *Creativity research journal* 18(3):391–404.

Davis, N.; Hsiao, C.-P.; Popova, Y.; and Magerko, B. 2015. An enactive model of creativity for computational collaboration and co-creation. In Zagalo, N., and Branco, P., eds., *Creativity in the Digital Age*. London: Springer. 109–133.

Elam, J. J., and Mead, M. 1990. Can software influence creativity? *Information Systems Research* 1(1):1–22.

Flaherty, A. W. 2005. Frontotemporal and dopaminergic control of idea generation and creative drive. *Journal of Comparative Neurology* 493(1):147–153.

Forgionne, G., and Newman, J. 2007. An experiment on the effectiveness of creativity enhancing decision-making support systems. *Decision Support Systems* 42(4):2126–2136.

Fujii, T.; Orihara, R.; Sei, Y.; Tahara, Y.; and Ohsuga, A. 2020. Generating cooking recipes from cooking videos using deep learning considering previous process with video encoding. In *Proc. of the 3rd Int. Conf. on Applications of Intelligent Systems*, APPIS 2020. New York, NY, USA: ACM.

Gero, K. I.; Liu, V.; and Chilton, L. 2022. Sparks: Inspiration for science writing using language models. In *Prof. of the 2022 ACM Designing Interactive Systems Conference*, DIS '22, 1002–1019. New York, NY, USA: ACM.

Gillies, M.; Fiebrink, R.; Tanaka, A.; Garcia, J.; Bevilacqua, F.; Heloir, A.; Nunnari, F.; Mackay, W.; Amershi, S.; Lee, B.; et al. 2016. Human-centred machine learning. In *Ext. Abstr. of the 2016 CHI Conf. on Human Factors in Computing Systems*, 3558–3565.

Gim, M.; Choi, D.; Maruyama, K.; Choi, J.; Kim, H.; Park, D.; and Kang, J. 2022. Recipemind: Guiding ingredient choices from food pairing to recipe completion using cascaded set transformer. In *Proc. of the 31st ACM Int. Conf. on Information & Knowledge Management*, CIKM '22, 3092–3102. New York, NY, USA: ACM.

Girotra, K.; Terwiesch, C.; and Ulrich, K. T. 2010. Idea generation and the quality of the best idea. *Management science* 56(4):591–605.

Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; and Holzinger, A. 2018. Explainable ai: the new 42? In *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 Int. Cross-Domain Conf., CD-MAKE 2018, Proceedings 2*, 295–303. Springer.

Goldschmidt, G. 2016. Linkographic evidence for concurrent divergent and convergent thinking in creative design. *Creativity research journal* 28(2):115–122.

Isaku, T., and Iba, T. 2015. Creative cocooking patterns: A pattern language for creative collaborative cooking. In *Proc. of the 20th European Conf. on Pattern Languages of Programs*, EuroPLoP '15, 1–17. New York, NY, USA: ACM.

Jenkin, T. A.; Chan, Y. E.; Skillicorn, D. B.; and Rogers, K. W. 2013. Individual exploration, sensemaking, and innovation: A design for the discovery of novel information: A design for the discovery of novel information. *Decision Sciences* 44(6):1021–1057.

Kamar, E. 2016. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *Proc. of the 25th Int. Joint Conf. on Artificial Intelligence*, 4070–4073.

Kato, F., and Hasegawa, S. 2013. Interactive cooking simulator: Showing food ingredients appearance changes in frying pan cooking -. In *Proc. of the 5th International Workshop on Multimedia for Cooking & Eating Activities*, CEA '13, 33–38. New York, NY, USA: ACM.

Kaufman, J. C., and Sternberg, R. J. 2010. *The Cambridge handbook of creativity*. Cambridge University Press.

Kletke, M. G.; Mackay, J. M.; Barr, S. H.; and Jones, B. 2001. Creativity in the organization: the role of individual creative problem solving and computer support. *International Journal of Human-Computer Studies* 55(3):217–237.

Kohn, N. W.; Paulus, P. B.; and Korde, R. M. 2011. Conceptual combinations and subsequent creativity. *Creativity Research Journal* 23(3):203–210.

Kuznetsov, S.; Rodriguez Vega, A.; and Long, E. 2022. A study of solar cooking: Exploring climate-resilient food preparation and opportunities for hci. In *Proc. of the 2022 CHI Conf. on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: ACM.

Larey, T. S. 1994. *Convergent and divergent thinking, group composition, and creativity in brainstorming groups*. The University of Texas at Arlington.

Lin, Y.; Guo, J.; Chen, Y.; Yao, C.; and Ying, F. 2020. It is your turn: Collaborative ideation with a co-creative robot through sketch. In *Proc. of the 2020 CHI Conf. on Human*

Factors in Computing Systems, CHI '20, 1–14. New York, NY, USA: ACM.

Marakas, G. M., and Elam, J. J. 1997. Creativity enhancement in problem solving: Through software or process? *Management Science* 43(8):1136–1146.

Medeiros, K. E.; Steele, L. M.; Watts, L. L.; and Mumford, M. D. 2018. Timing is everything: Examining the role of constraints throughout the creative process. *Psychology of Aesthetics, Creativity, and the Arts* 12(4):471–488.

Mobley, M. I.; Doares, L. M.; and Mumford, M. D. 1992. Process analytic models of creative capacities: Evidence for the combination and reorganization process. *Creativity Research Journal* 5(2):125–155.

Mumford, M. D.; Baughman, W. A.; Supinski, E. P.; and Maher, M. A. 1996. Process-based measures of creative problem-solving skills: Ii. information encoding. *Creativity Research Journal* 9(1):77–88.

Naritomi, S., and Yanai, K. 2021. 3d mesh reconstruction of foods from a single image. In *Proc. of the 3rd Workshop on AIxFood*, AI & Food'21, 7–11. New York, NY, USA: ACM.

Plucker, J. A. 2004. Generalization of creativity across domains: Examination of the method effect hypothesis. *The Journal of Creative Behavior* 38(1):1–12.

Policastro, E., and Gardner, H. 1999. From case studies to robust generalizations: An approach to the study of creativity. In Sternberg, J., ed., *Handbook of creativity*. Cambridge University Press. 213–225.

Punpongsanon, P.; Miyatake, Y.; Iwai, D.; and Sato, K. 2022. Demonstration of interiqr: Unobtrusive edible tags using food 3d printing. In *Adjunct Proc. of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22 Adjunct. New York, NY, USA: ACM.

Rietzschel, E. F.; Slijkhuis, J. M.; and Van Yperen, N. W. 2014. Task structure, need for structure, and creativity: Task structure and creativity. *European Journal of Social Psychology* 44(4):386–399.

Runco, M. A., and Acar, S. 2012. Divergent thinking as an indicator of creative potential. *Creativity research journal* 24(1):66–75.

Sauvé, K.; Bakker, S.; and Houben, S. 2020. Econundrum: Visualizing the climate impact of dietary choice through a shared data sculpture. In *Proc. of the 2020 ACM Designing Interactive Systems Conference*, DIS '20, 1287–1300. New York, NY, USA: ACM.

Sawyer, R. K. 2011. *Explaining creativity: The science of human innovation*. Oxford university press.

Schleith, J.; Norkute, M.; Mikhail, M.; and Tsar, D. 2022. Cognitive strategy prompts: Creativity triggers for human centered ai opportunity detection. In *Creativity and Cognition*, C&C '22, 29–37. New York, NY, USA: ACM.

Sherer, M.; Maddux, J. E.; Mercandante, B.; Prentice-Dunn, S.; Jacobs, B.; and Rogers, R. W. 1982. The self-efficacy scale: Construction and validation. *Psychological reports* 51(2):663–671.

Shneiderman, B. 2002. Creativity support tools. *Commun. ACM* 45(10):116–120.

Simonton, D. K. 2015. On praising convergent thinking: Creativity as blind variation and selective retention. *Creativity Research Journal* 27(3):262–270.

Sternberg, R. J., and Lubart, T. I. 1999. The concept of creativity: Prospects and paradigms. *Handbook of creativity* 1(3-15).

Svensson, M.; Höök, K.; and Cöster, R. 2005. Designing and evaluating kalas: A social navigation system for food recipes. *ACM Trans. Comput.-Hum. Interact.* 12(3):374–400.

Toubia, O. 2006. Idea generation, creativity, and incentives. *Marketing science* 25(5):411–425.

Treffinger, D. J.; Isaksen, S. G.; and Stead-Dorval, K. B. 2006. *Creative problem solving: An introduction*. Prufrock Press Inc.

Urban Davis, J.; Anderson, F.; Stroetzel, M.; Grossman, T.; and Fitzmaurice, G. 2021. Designing co-creative ai for virtual environments. In *Creativity and Cognition*, C&C '21. New York, NY, USA: ACM.

Wang, K., and Nickerson, J. V. 2017. A literature review on individual creativity support systems. *Computers in Human Behavior* 74:139–151.

Wang, H., and Ohsawa, Y. 2013. Idea discovery: A scenario-based systematic approach for decision making in market innovation. *Expert Systems with Applications* 40(2):429–438.

West, J.; Fusari, G.; Raby, E.; Alwani, R.; Meldaikyte, G.; Wojdecka, A.; and Matthews, E. 2018. Developing the double diamond process for implementation. In Barron, D., and Seemann, K., eds., *Proc. of the 4th Int. Conf. on Design4Health*, volume 1, 310–312.

Wu, T.; Jiang, E.; Donsbach, A.; Gray, J.; Molina, A.; Terry, M.; and Cai, C. J. 2022. Promptchainer: Chaining large language model prompts through visual programming. In *Ext. Abstr. of the 2022 CHI Conf. on Human Factors in Computing Systems*, CHI EA '22. New York, NY, USA: ACM.

Wu, T.; Terry, M.; and Cai, C. J. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proc. of the 2022 CHI Conf. on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: ACM.

Yang, Q.; Steinfeld, A.; and Zimmerman, J. 2019. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proc. of the 2019 CHI Conf. on Human Factors in Computing Systems*, CHI '19, 1–11. New York, NY, USA: ACM.

Yoneda, Y., and Nadamoto, A. 2018. Knack for cooking extraction from user generated recipe sites. In *Proc. of the 20th Int. Conf. on Information Integration and Web-Based Applications & Services*, iiWAS2018, 134–137. New York, NY, USA: ACM.

Interdisciplinary Methods in Computational Creativity: How Human Variables Shape Human-Inspired AI Research

Nadia M. Ady[*1,2] **and Faun Rice**[*3]
[1]Department of Computing Science, University of Alberta
[2]Alberta Machine Intelligence Institute (Amii)
[3]Independent Researcher
nmady@ualberta.ca, faun.rice@gmail.com

Abstract

The word *creativity* originally described a concept from human psychology, but in the realm of computational creativity (CC), it has become much more. The question of what creativity means when it is part of a computational system might be considered core to CC. Pinning down the meaning of creativity, and concepts like it, becomes salient when researchers port concepts from human psychology to computation, a widespread practice extending beyond CC into artificial intelligence (AI). Yet, the human processes shaping human-inspired computational systems have been little investigated. In this paper, we question *which* human literatures (social sciences, psychology, neuroscience) enter AI scholarship and *how* they are translated at the port of entry. This study is based on 22 in-depth, semi-structured interviews, primarily with human-inspired AI researchers, half of whom focus on creativity as a major research area. This paper focuses on findings most relevant to CC. We suggest that *which* human literature enters AI bears greater scrutiny because ideas may become disconnected from context in their home discipline. Accordingly, we recommend that CC researchers document the decisions and context of their practices, particularly those practices formalizing human concepts for machines. Publishing reflexive commentary on human elements in CC and AI would provide a useful record and permit greater dialogue with other disciplines.

Introduction

Computational creativity (CC) is informed by many human literatures, including psychology, sociology, cognitive science, and philosophy (Ackerman et al., 2017, p. 11; McGregor, Wiggins, and Purver, 2014). There is a long history of reflection on the relationship between CC's parent, AI, with other disciplines (Newell, 1970) which continues today (Lieto and Radicioni, 2016; MacPherson et al., 2021; Cassenti, Veksler, and Ritter, 2022). Social sciences also offer relevant commentary: Science, Technology, and Society (STS) is concerned with how scientific methods produce knowledge and shape the world, calling attention to the human processes inherent in scientific work using a broad methodological toolkit (Jasanoff, 2013; Lippert and Mewes, 2021; Law, 2004; Suchman and Trigg, 1993). In alignment with these conversations, our project explores the human processes involved when researchers draw inspiration from concepts from human psychology for computational systems. In this paper, we present early findings centred on CC. This work responds to calls to articulate the "methodological and conceptual barriers … [which] confront attempts to work across disciplinary boundaries" (MacLeod, 2018, p. 697).

Our dataset is 22 in-depth, semi-structured interviews with CC and AI researchers working closely with concepts from human psychology (see Methodology). For 11 interviewees, the concept of creativity is a key thread in their research: the other 11 engaged with concepts such as curiosity, forgetting, or mental time travel. We use "human-inspired" as shorthand for this heterogeneous group throughout the paper, and transcripts from non-CC participants refine our understanding of each finding, though our focus here is on CC. We build on existing scholarship by suggesting that human and social factors impact *which* human literature enters AI and *how* it is translated for computation at its port of entry. Further, we suggest that human and social processes in CC are productive areas of inquiry, and that qualitative methods offer fruitful ways of exploring these topics, in agreement with scholars like Pérez y Pérez and Ackerman (2020). In demonstration, we outline two phenomena related to the challenges of interdisciplinary work, followed by an example of intellectual influence on human-inspired AI that emerged from qualitative interviews.

Methodology

This study has used a grounded theory approach to conception, data collection, and analysis. Aligned with grounded theory methodologies, we began with a broad interest rather than a hypothesis (Qureshi and Ünlü, 2020); prioritized inductive findings from primary qualitative research (Glaser and Strauss, 1967); and participated collaboratively in transcription, line-by-line coding, memoing, focused coding, and forming early-stage conceptual categories (Wiener, 2007, p. 301; Charmaz, 2014).

We began with purposive sampling of human-inspired CC and AI researchers. We used interviewees' publications to assess their relevance to study aims, and proceeded via snowball sampling. In one-hour long semi-structured interviews, we asked participants how they defined the hu-

[*]Both authors contributed equally to this work.

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

man concepts they worked with, what types of literature and personal experience had shaped their definitions, and what challenges or successes they encountered in translating their concepts for machines.

Of our 22 participants, six used she/her pronouns, and all worked in North America or Europe; improved gender and regional diversity are goals of this study as we continue data collection. Participants included five employees of private sector AI firms, three PhD students, one postdoctoral researcher, four pre-tenure professors, and nine post-tenure professors. Academic interviewees were primarily in computational- or psychology-related departments. We use the convention P# to anonymize participants.

Grounded theory methodology espouses simultaneous data collection and analysis, suggesting that questions raised by early rounds of analysis should be pursued further in subsequent data collection as a form of theoretical sampling, alongside further review of relevant literature (Charmaz, 2014). Early dissemination through this paper allows us to incorporate diverse feedback into the study (Green et al., 2007, p. 489). Accordingly, we will further develop this project by taking cues from readers in the CC community, including additional data collection and deeper exploration of themes introduced in this paper. This project was approved by University of Alberta Research Ethics Board 1 (ID: Pro00109111).

How do Ideas from Human Literatures Enter CC and AI?

The difficulty of reading broadly for interdisciplinary research was a core theme in interviews. Several interviewees felt that "understanding what is going on in all [the] different, relevant fields" was "one of [their] biggest challenges as researchers," (P22). They often relied on "serendipity," (P22) or popular culture: "I'm probably more likely to learn from a New York Times article profiling the day in the life of an artist than I am to actually read art history books" (P15). Others began conversations at conferences or across departments, or employed strategies like citation chaining to make headway. The precise language (jargon) required for high levels of rigor is known to sometimes prevent access by readers external to a discipline, meaning that this challenge may come with the territory of interdisciplinary work (Callaos and Horne, 2013, pp. 23-24; Daniel et al., 2022 pp. 8-9; MacLeod, 2018, p. 707).

As a result of this challenge, keeping up with ongoing debate and discussion in the human literatures may become deprioritized once an idea has gained traction in CC or AI. In some cases, knowledge about ideas' origins may be lost. P6 offered the example of catastrophic forgetting in machine learning, connecting it to the psychological term "retroactive inhibition," and contending that the usage in machine learnin was initially congruent with the usage in psychology but became increasingly misaligned. Authors became less aware of the origins of the ideas they were citing over time. P6 sees a "strong disconnect" as a consequence: "the concept of forgetting as it appears in the psychological literature is much more broad and diverse than forgetting within artificial neural networks" (P6).

The first time a concept is ported into AI from human literatures, it may be from a seminal scholar doing important translation work for their field. Margaret Boden's model of creativity as novelty, value, and surprise (2004, p. 1), for example, was described by many as a "huge service" which "put forward ideas … that had been around long before her work in the 1970s" but introduced them to cognitive and computer science (P12). Nevertheless, the fruits of translation work may still lose contact with ongoing discussions in other fields, and even ossify in CC or AI: for example, one interviewee suggested that citing Margaret Boden had simply become part of the brand of CC (P21), and another recalled defaulting to citing her work when rebuked by "senior academics" for not "being specific in [their] definition" (P3).

Similarly, scholars in other disciplines may pave the way for cross-disciplinary translation by synthesizing a topic in a popular non-fiction book or textbook. Such works can acquire a kind of virality in AI and its subfields. For example, multiple interviewees mentioned the works of Csikszentmihalyi (e.g. *Flow*, 1990) and Tomasello (*Becoming Human*, 2019). P11 described *Becoming Human* (2019) as "making the rounds among AI–psychology academics. It's basically about what differentiates children from primates … spanning development, psychology, primatology and others where it really hones in on what capabilities humans have … And I just think that there are people that ask those questions and [Tomasello] presents them very clearly." Textbooks and popular non-fiction offer essential knowledge translation, but come with constraints (e.g., editorial standards sometimes discourage authors from citing in-text; authors are offering a broad synthesis of ideas) that do not provide a full understanding of ongoing conversations in the authors' home disciplines (Callaos and Horne, 2013, pp. 23-24).

We do not intend to undervalue influential cross-disciplinary contributions: interviewees saw such work as crucial and valuable, but simply warned of "tunnel vision" (P9) or "misalignment" (P6). When a single scholar becomes 'the person' to cite, it may permit a disjuncture between AI and ongoing debates in psychology or other disciplines. Other scholars have raised concerns about "herd mentality" (Rekdal, 2014, p. 570) or maintenance of the status quo (Dworkin, Zurn, and Bassett, 2020, p. 890) in contemporary citation practices. At the same time, citations help us position thinkers "as the authors, authorities and originators" and remember and acknowledge the genealogy of a field (Liu, 2021, p. 215) and our debts "to those who came before" (Ahmed, 2017, p. 15). Citing Boden in particular, might be considered a feminist practice given the disproportionate number of men in AI, reminding the reader that CC is indebted to a woman scholar.

Ironically, feeling that one has to adequately acknowledge an idea's disciplinary context may deter scholars from taking inventive paths to new knowledge–for example, one graduate student interviewee commented, "that's the difficulty with interdisciplinary work. You're never going to be just one or the other. You can't be the best at either field by devoting yourself to both, right?" (P1). **Yet, we would sug-**

gest that researchers should not be afraid to read broadly despite knowing that some disciplinary context will be lost. While key pieces of knowledge translation can generate valuable new lines of thinking, it is also important to recognize that no one source can explain a whole field. Instead, researchers can think through the effects of which sources from human literatures are influencing their work. Reflecting on the influence of different ideas and disciplines might also allow CC researchers to identify gaps and opportunities for future research: for example, it is possible that work in English is more likely to make its way into CC. Similarly, ideas about creativity from disciplines beyond psychology and neuroscience (e.g., sociology, anthropology, philosophy) may fall outside of typical reading lists for historical and institutional reasons.

'An Interpretation Job': Articulating Hidden Methodological Decisions

Interviewees described a second challenge related to translating ideas from human literatures to CC or AI. Many found concepts in "the human literature [to be] not well defined… more of a nice metaphor" (P8), "open ended and ambiguous" (P22), or "extremely general" (P14). For example, interviewees described digging through work by Jean Piaget, Lev Vygotsky, or Daniel Berlyne seeking clarity on theoretical terminology. Others observed that many concepts might not yet have direct parallels in AI: "Appraisal theory of motivation, for instance, that's very much connecting motivation with emotion or affect. And what does that even correspond to in AI? Now that's quite a challenge. Is it even worth going down that route if we don't have any corresponding elements for that in the computational domain?" (P22).

As a result, some researchers attribute their choice of definition to ease of evaluation or formalization. Such decisions include discarding some definitions, like steering away from "*H-creativity* [historical creativity], which is only theoretical, because how do you even measure [the idea] that no one in human history has ever had this thought before" (P3). Alternatively, P16 described looking through competing definitions in psychology and landing on one "easily translatable into reinforcement learning." Finally, ease of evaluation might present an opportunity to contribute: P22 expressed enthusiasm for "modeling very minimal creativity because [they] think it is more amenable to measurement."

While interviewees described some ideas from human literatures as more amenable to evaluation or formalization than others, processes of translation across that spectrum involved individual choices. We asked how interviewees "translated" definitions for computational systems, and one countered, "it's not so much of a translation problem as that the first definition is blurry. It's more of an interpretation type of job" (P16). P14 described a similar process: "to do this translation of these psychological ideas, especially when they are ambiguous like most of them are, this process is what really makes the difference. It's not simple to do… And there are a lot of things [in this process] that are so important for the development of knowledge and insight." Articulating choices made during translation of psychological ideas for computer science not only helps future "translators," it also tracks the changes in meaning that concepts may undergo during this process.

Accordingly, **there are two stages where researchers might record decisions they make about using ideas from human literatures: first, at the point of selecting or discarding particular definitions, and second, during interpretation.** The use of reflexive (self-aware, positional) description of one's own research processes as data is well established in the social sciences and STS (e.g., Soler et al., 2014, pp. 12-13), and some CC researchers have begun to adopt similar methods (e.g., Pérez y Pérez and Ackerman, 2020). Fiske (2020) offers an exemplar of how reflexive description can contribute to a field. They describe the process of becoming interested in the concept of feeling "*moved*" and looking for explication in the literature and primary research. By walking through human elements in the research process and the way that social and institutional factors shaped conclusions, Fiske offers novel commentary on their discipline's methods while describing their own findings. Ultimately, "polyglotism of [their] research group … helped protect [them] from the lexical fallacy of conflating the usage of a vernacular lexeme–say, *be moved*–with the features of a mental state" (2020, p. 97). In a similar way, **improved legitimacy of reflexive data in computer science and human-inspired AI may articulate hidden methodological decision making for researchers in AI and also provoke productive novel discussions.**

Implicit Levels of Analysis

We have suggested thus far that reflecting on the processes of reading, selecting, and interpreting ideas from other disciplines offers an opportunity to CC researchers working with concepts from human literatures. In this section, we close with an example of an influential idea from human literatures that was present implicitly or explicitly in many of our interviews. Improved awareness of this idea's origins might help CC researchers, and more broadly AI researchers, articulate methodological beliefs and decisions.

While describing the concepts they work with, many interviewees invoked the idea of "levels" of explanation, description, reduction, understanding, abstraction, or analysis. There is a long history of considering human and machine minds in terms of levels (see for example, Putnam, 1975; Marr and Poggio, 1976; Poggio, 2012; Schouten and Looren de Jong, 2007; MacDougall-Shackleton, 2011). In psychology, the use of "levels" is tied to the question of whether higher level constructs in the "mind" can be reduced to physical processes in the brain, beginning with early psychoneural reductionists (Feigl, 1958; Place, 1956; Smart, 1959 in Schouten and Looren de Jong, 2007, p. 6). This use is therefore deeply tied to the history of cognitive science and AI. Furthermore, such levels are used more broadly in philosophy of science to consider the question of reductionism across disciplines and share a common intellectual history if not identical terminology (Schouten and Looren de Jong, 2007).

Interestingly, while some interviewees explicitly referenced levels of analysis and cited the computational version described by Marr and Poggio (1982, 1976) either in conversation or in scholarly work (e.g., Wiggins, 2020, p. 2; Dasgupta, 2019, pp. 2, 55), many used the language of levels implicitly, and perhaps unconsciously: "psychologists seem to be willing to treat underlying learning as a mystery, which I think is appropriate, and have their models of it. And that's what I think we're doing [in AI], even though as researchers we don't often realize it" (P17).

MacDougall-Shackleton (2011), writing on diverse uses of levels of analysis in studies of human and animal behaviour, characterizes the essential difference between levels of analysis as between proximate mechanisms vs. ultimate functions. Similarly, many CC interviewees articulated a difference between evaluating creativity based only on an output, like a musical composition, or evaluating creativity based on some aspect of process or mechanism (Jordanous, 2016, p. 1).

Meta-commentaries on levels of analysis have suggested that inquiry at any level (mechanism or outcome or any variation thereof) is a valid approach but should be specified clearly (Schouten and Looren de Jong, 2007; MacDougall-Shackleton, 2011). They further identify many false debates between process and product (ibid.). Several interviewees either engaged in a process/product debate or acknowledged its presence in CC: for example, P4 described feeling that something was missing when "[another scholar] came up with a set of criteria for how to evaluate the [creativity of the] output of a system. I couldn't quite accept that because it didn't take any notice of the process that the system was using." Another articulated the importance of distinguishing between *ex post* and *ex ante* definitions in CC, where *ex post* creativity means that creativity is realized when, once achieved, you "can't tell the difference any more" between the type of mind responsible for an outcome (P12)—expressed by Putnam (1975, p. 291), "we could be made of Swiss cheese and it wouldn't matter." This conversation is also present in CC literature, with Pease and Colton (2011) distinguishing between "(i) judgements which determine whether an idea or artefact is valuable, and (ii) judgements to determine whether a system is acting creatively" (p. 72, cf. Wiggins, 2021; Hodson, 2017).

This example of the levels of analysis lends support to our hypothesis that reflection on the intellectual history of influential ideas from human literatures in CC could lend additional clarity, in this case, to debates about definitions of creativity. Researchers engaging in reflexive commentary on their own work, as suggested in the previous section, might consider documenting the level of analysis they are working with (and whether they intend to remain on that level or take an intentionally integrationist approach; see also Poggio, 2012). With respect to our project, we have found that levels of description are implicit in our ability to articulate our scope. This project is concerned with where different researchers' definitions/understandings of concepts like creativity come from and how they influence subsequent thinking; at what levels of analysis are our project's scope-defining concepts, like creativity, understood? Sharpening our terminology is an ongoing project for us, so we recognize its difficulty, but we also found that levels of analysis continue to help us understand how interviewees formed, implemented, and evaluated their concepts. Indeed, one interviewee explicitly expressed that it would be "really valuable" for AI researchers to "have a little bit more understanding of things like Marr's levels of analysis" (P10).

Conclusion and Future Work

This paper argues that sharper attention to human elements in research can help generate novel perspectives for CC, and human-inspired AI research more generally. It explores the challenges of interdisciplinary research, suggesting that attention to translation work (including what is being ported across disciplines and how) is beneficial to CC. However, this paper has several limitations: primarily, we present early stage findings. As this project proceeds, we will refine our discussions through additional data collection, seeking theoretical saturation. The term *theoretical saturation* refers to the point at which gathering more data no longer yields further theoretical insights (Bryant and Charmaz, 2007, p. 611). For example, we expect that salient new advances like the release of OpenAI's ChatGPT will contribute to further development of this project, as data collection for this study primarily took place before November 2022. Furthermore, we intend to expand the interview sample to seek psychologists' perspectives on ideas from human literatures that seem to be highly influential in AI and on whether the representations of those ideas in AI are representative of current discussions in psychology.

Finally, future work will expand the diversity of experience represented by interviewees. In addition to being researchers, interviewees were often artists and had biographical elements informing their work. Rich individual histories influence what researchers see as valuable and what brings them satisfaction or joy, and therefore what is worth studying. Personal narratives and their role in scientific discovery were raised by interviewees (for example, P7 told a story of drawing inspiration from losing a game of 'memory,' or 'concentration,' to their child) and these narratives clearly play a role in methods and approach. For interviewees in computational creativity, the influence of personal values often included a commitment to uplift rather than replace human creativity (P22, for example, discussed avoiding imitating "*Big-C* creativity" in part because of the potentially severe ethical consequences of doing so). This paper has touched on biography and personal choice, and future work may further develop these themes.

While this paper has sought to outline some of the human influences on scientific research progress, it does not claim that they can or should be eliminated: rather, that elucidating them will help prompt clearer reflection how human-inspired AI is shaped. As P13 commented, "I've learned early on that the best research comes from our lived experience and intuition about the world, and once you have some hypotheses then you apply the scientific method and do things properly, but it's motivated by our own experiences and that's where I see the best work getting done. I don't think we can research something that we don't live."

Acknowledgements

We are grateful to numerous contributors, not limited to M. Ackerman, D. Brown, W. Carvalho, J. Casey, C. Colas, G. Cottrell, I. Dasgupta, J. Droppo, R. French, K. Gordon, C. Guckelsberger, K. Gupta, M. Guzdial, A. Jordanous, A. Lampinen, K. Mathewson, A. McLafferty, A. Parrish, R. Pérez y Pérez, P. M. Pilarski, R. Saunders, R. Shariff, W. Tanenbaum, J. Virani, G. Wiggins, and M. Wilson. NMA was funded by the Canada CIFAR AI Chairs program and the Alberta Machine Intelligence Institute (Amii).

References

Ackerman, M.; Goel, A.; Johnson, C. G.; Jordanous, A.; León, C.; Pérez y Pérez, R.; Toivonen, H.; and Ventura, D. 2017. Teaching computational creativity. In *Proc. Int. Conf. Computational Creativity (ICCC'17)*, 9–16.

Ahmed, S. 2017. *Living a Feminist Life*. Duke University Press.

Boden, M. A. 2004. *The Creative Mind*. London: Routledge.

Bryant, A., and Charmaz, K. 2007. *The SAGE Handbook of Grounded Theory*. SAGE Publications.

Callaos, N., and Horne, J. 2013. Interdisciplinary communication. *J. Syst. Cybern. Inform.* 11(9):23–31.

Cassenti, D. N.; Veksler, V. D.; and Ritter, F. E. 2022. Editor's review and introduction: Cognition-inspired artificial intelligence. *Top. Cogn. Sci.* 14(4):652–664.

Charmaz, K. 2014. *Constructing Grounded Theory*. Introducing Qualitative Methods. SAGE Publications.

Csikszentmihalyi, M. 1990. *Flow*. New York: Harper & Row.

Daniel, K. L.; McConnell, M.; Schuchardt, A.; and Peffer, M. E. 2022. Challenges facing interdisciplinary researchers: Findings from a professional development workshop. *PLOS ONE* 17(4):1–16.

Dasgupta, I. 2019. *Algorithmic approaches to ecological rationality in humans and machines*. Ph.D. Dissertation, Harvard University.

Dworkin, J.; Zurn, P.; and Bassett, D. S. 2020. (In)citing action to realize an equitable future. *Neuron* 106(6):890–894.

Fiske, A. P. 2020. The lexical fallacy in emotion research: Mistaking vernacular words for psychological entities. *Psychol. Rev.* 127(1):95–113.

Glaser, B. G., and Strauss, A. L. 1967. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine de Gruyter.

Green, D. O.; Creswell, J. W.; Shope, R. J.; and Clark, V. L. P. 2007. Grounded theory and racial/ethnic diversity. In Bryant, A., and Charmaz, K., eds., *The SAGE Handbook of Grounded Theory*. SAGE Publications. 472–492.

Hodson, J. 2017. The Creative Machine. In *Proc. Int. Conf. Computational Creativity (ICCC'17)*, 143–150.

Jasanoff, S. 2013. Fields and fallows: A political history of STS. In *Interdisciplinarity*. Routledge. 99–118.

Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connect. Sci.* 28(2):194–216.

Law, J. 2004. *After Method*. Routledge.

Lieto, A., and Radicioni, D. P. 2016. From human to artificial cognition and back: New perspectives on cognitively inspired AI systems. *Cogn. Syst. Res.* 39:1–3.

Lippert, I., and Mewes, J. S. 2021. Data, Methods and Writing: Methodographies of STS Ethnographic Collaboration in Practice. *Sci. Technol. Stud.* 34(3):2–16. Number: 3.

Liu, X. 2021. The use/less citations in feminist research. *Aust. Fem. Stud.* 36(108):212–221.

MacDougall-Shackleton, S. A. 2011. The levels of analysis revisited. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 366(1574):2076–2085.

MacLeod, M. 2018. What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice. *Synthese* 195(2):697–720.

MacPherson, T.; Churchland, A.; Sejnowski, T.; DiCarlo, J.; Kamitani, Y.; Takahashi, H.; and Hikida, T. 2021. Natural and artificial intelligence: A brief introduction to the interplay between AI and neuroscience research. *Neural Netw.* 144:603–613.

Marr, D., and Poggio, T. 1976. From understanding computation to understanding neural circuitry. Tech. Rep. AIM-357, MIT.

Marr, D. 1982. *Vision*. W. H. Freeman and Company.

McGregor, S.; Wiggins, G. A.; and Purver, M. 2014. Computational creativity: A philosophical approach, and an approach to philosophy. In *Proc. Int. Conf. Computational Creativity (ICCC'14)*, 254–262.

Newell, A. 1970. Remarks on the relationship between artificial intelligence and cognitive psychology. In Banerji, R. B., and Mesarovic, M. D., eds., *Theoretical Approaches to Non-Numerical Problem Solving*, 363–400. Springer.

Pease, A., and Colton, S. 2011. Computational creativity theory: Inspirations behind the FACE and the IDEA models. In *Proc. Int. Conf. Computational Creativity (ICCC'11)*, 72–77.

Poggio, T. 2012. The *Levels of Understanding* framework, revised. *Perception* 41(9):1017–1023.

Putnam, H. 1975. Philosophy and our mental life. In *Philosophical Papers*, volume 2. Cambridge University Press. 291–303.

Pérez y Pérez, R., and Ackerman, M. 2020. Towards a Methodology for Field Work in Computational Creativity. *New Gener. Comput.* 38(4):713–737.

Qureshi, H. A., and Ünlü, Z. 2020. Beyond the paradigm conflicts: A four-step coding instrument for grounded theory. *Int. J. Qual. Methods* 19.

Rekdal, O. B. 2014. Academic citation practice: A sinking sheep? *portal* 14(4):567–585.

Schouten, M., and Looren de Jong, H. 2007. Mind matters: The roots of reductionism. In Schouten, M., and Looren de Jong, H., eds., *The Matter of the Mind*. Blackwell. 1–28.

Soler, L.; Zwart, S.; Lynch, M.; and Israel-Jost, V. 2014. Introduction. In *Science after the Practice Turn in the Philosophy, History, and Social Studies of Science*. Routledge. 1–43.

Suchman, L. A., and Trigg, R. H. 1993. Artificial intelligence as craftwork. In Chaiklin, S., and Lave, J., eds., *Understanding Practice*. Cambridge University Press. 144–178.

Tomasello, M. 2019. *Becoming Human*. Harvard University Press.

Wiener, C. 2007. Making teams work in conducting grounded theory. In Bryant, A., and Charmaz, K., eds., *The SAGE Handbook of Grounded Theory*. SAGE Publications. 293–310.

Wiggins, G. A. 2020. Creativity, information, and consciousness: The information dynamics of thinking. *Phys. Life Rev.* 34-35:1–39.

Wiggins, G. A. 2021. Computational creativity and consciousness: Framing, fiction and fraud. In *Proc. Int. Conf. Computational Creativity (ICCC'21)*, 183–191.

Stories Stay, Lessons Leave: Principles on AI Art from Photography

Alayt Issak
College of Arts, Media and Design
Northeastern University
Boston, MA 02120 USA
issak.a@northeastern.edu

Abstract

The recent popularity of Artificial Intelligence (AI) within the arts is a phenomenon that maps a continuation of the historical advances in technology. From the insurgence of calligraphy with the introduction of the letterpress, to the photographs shared space in painting, technology has entered the arts with its most recent immersion via AI. In this work, we acknowledge that AI carries affordances to art that is not present in other medium but equally raise questions on what defines its aesthetics as a notion of art. As such, we draw parallels from conceptual photography's historical past to synthesize principles of interaction, along with instantiated examples, to carry AI Art toward an aesthetic discipline.

Introduction

Art, according to contemporary definitions and schools of thought, can be defined in a traditional, aesthetic, or institutional practice. In our focus on aesthetic principles, we highlight artworks derived through conceptual experience that fall into question when AI is integrated as a medium (Manovich and Arielli 2023) - for which we define AI Art. As such, we query how framing the discussion on the integration of AI as a medium can open creative affordances that would not be available otherwise.

Much often, questions on AI Art fall under a critique of the medium, such as its utility for regurgitating data in interesting ways, which in some cases has led to hesitations to even attribute art that is created with AI (Mikalonytė and Kneer 2022). Yet, this phenomenon is not novel, as warranting a closer look into art history, finds that photography was equally criticized, in its case, for mechanically plastering reflections of reality that do not compare to the labor of a painter.

During photography's contested incline in the 20th century, seminal notable photographers shared their say on the medium. Edward Weston claimed photography had "opened blinds to a new world vision" whereas Paul Strand was indifferent to the question of whether photography is an art (Sontag 2001, p. 96). However, the conceptual notion of photographic style in the following decades carried a desire to let the medium be without comparison to its compatriot arts, at least for those who subjected themselves to its use.

Soon enough, with movements such as Pictorialism (Sternberger 2001), photography was uncontested within the arts, and although this could be a factor of time, this phenomenon equally emerged through shifts of interaction that opened modes of aesthetics that were not available before.

Approach

Taking lessons from photography without rehearsing history, we find that framing AI Art towards its own stronghold can unveil aesthetic affordances that are easily adaptable. To begin by opening contemporary doors, it can be understood that "computers do not create art, people using computers create art" (Hertzmann 2018, p. 2). This allows one to view AI as a medium that is to decipher its mode of interaction.

To follow suit, with AI Art having greatly accelerated since 2014, endeavors on various fronts begin to guide its lens. These include technical illustrations (Shan et al. 2023; Zammit, Liapis, and Yannakakis 2022), synopsis on discourse (Newton and Dhole 2023; Issak and Varshney 2022), literature on the novelty of encoding artistic input as an abstract multi-dimensional space of image representations (Cetinic and She 2021, p. 9), and introductions of ethics (Divakaran, Sridhar, and Srinivasan 2023; Ventura and Gates 2018) and explainability (Bodily and Ventura 2018) as measures to guide system aesthetics in such pursuits.

However, to elucidate the mapping of AI Art in greater depth, we interweave parallels to conceptual photography through current anthology and engage in research through design (RtD) to devise our findings. In doing so, we develop three principles of interaction, namely in technique, utility, and phenomenology, coupled with examples for instantiation. Our principles are largely informed by the distinction between what is 'artistic' and 'aesthetic' from a pragmatic approach as we seek to overarchingly navigate AI Art toward an aesthetics discipline.

Principle 1: Technique creates a discipline

One may say that photography emerged as a discipline in its ability to transform from a tool to view reality, such as the initial Daguerreotypes (Kul-Want 2010, p. 105), into a discipline of its own in unveiling an extension to view the world, such as the minuscule details of everyday visuals. That is, by establishing a method that does not compare to its counterparts, such as capturing the fleeting moments of the human

Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)
ISBN: 978-989-54160-5-9

eye, be it the speed of a horse or the momentary sighting of a beautiful sunset.

Yet this does not suffice to deduce its triumph. A camera that simply allows one to press a button is different from one that takes the artist through the journey of the artwork. That was the objection raised by painters as they challenged the Daguerrotype's instantaneous mimicry of reality without the inherent experience in mind. Certainly, photography had its own process that undertook the artist through a journey. The preparation of the camera was an act of its own but, as elucidated in Walter Benjamin's 1931 essay "A Small History of Photography" (Kul-Want 2010, p. 110), the technique is what ultimately gave rise to its uncontested foothold.

Photographers may speak about their process of tuning on exposure times similar to how a painter may describe selecting different brushes, and a photographer's eye may now be tailored differently to capture what other disciplines would not. These factors lend questions that are specific for the discipline to decipher. In fact, when photography even began to emerge on its own, artists began to extract the tool from the process. Some examples of such extraction include the cyanotype method developed by Anna Atkins in 1842 to print a white negative on a blue (cyan) colored background (Lotzof 2018), the "printing out" method of 1891 for printing images without a darkroom, and photograms to recite "photography without a camera" as coined by Laszlo Moholy-Nagy in the early 20th century (Moholy-Nagy and Molderings 2010).

These examples recognize the importance of technique that may even supersede the tool as they resulted in "photography" as a discipline that is not based on the camera that was used to capture images, i.e. the medium that led to its contested beginning in the first place. It has established through its technique a world view that is now observed and carried out upon various mediums. With this realization, we translate one possibility in which AI Art may also open a worldview through its intrinsic abilities.

Examples of AI Art Technique

In this context, we employ AI's mechanism to view nuances of the everyday imaginations of our world. We highlight "imagination" as what we observe not what is visible to the eye, but rather what is invisible. In tangential efforts, these have been used for idea generation within the creative process (Smith et al. 2023), however, we frame our inquiry on the intangible ideation.

Title: Electrical Activity in Fungi

In the following implementation, we query the presence of electrical activity in Fungi, a line of research that has been explored for the possibility of fungal electric transmission. We base our prompts on the limitations to this query - the difficulty detecting the variable analysis of neural activity in Fungi (Dehshibi and Adamatzky 2021). We utilize Stable Diffusion, a high-resolution image-synthesis model, (Rombach et al. 2022) to illuminate our ideation in two detailed formats, which we arrive at upon prompting various iterations that signaled the posited inquiry.

In the nascent image the idea behind the prompt paints, what differentiates this seed from a photograph, is that it is not capturing the nuanced physical elements of fungi, but rather presents the idea as an amalgamation of what has been learned about this process. In fact, although this may fall under the fallacy to represent an actuality, which can be excused as a seedling of imagination, it is not a microscope or pure imagination, but somewhere in between. In doing so, Fig 1a displays what the "imagined data" looks like, whereas Fig 1b dives deeper into the space of possible imagery.

(a) Prompt 1: An intricate sci-fi VR 3D painting of electrical activity in Fungi showing the spiking activity of the mycelium networks with movement about mechanisms.

(b) Prompt 2: An intricate sci-fi VR 3D painting of electrical activity in Fungi showing the spiking activity of the mycelium networks **detecting** the activity about mechanisms.

Figure 1: Electrical Activity in Fungi: External and Internal

Principle 2: Transition from tool to instrument

In a pragmatic approach, John Dewey, in his book, *Art as Experience*, cites the distinction between the artistic and esthetic (Dewey 2005). In Chapter III, Having an Experience, Dewey elucidates the difference between what is 'artistic' (act of production) and 'esthetic' (perception and enjoyment), in that no experience of any sort is in unity unless it has esthetic quality, which occurs in the alternating relationship of doing and undergoing, and is joined by perception. This is a notion in art that entails esthetic experience as inherently connected with the experience of making.

To Dewey's point, 'aesthetics' (experience) is the judgment and 'artistic' (art) is the expression, and interpreting this definition finds place in transitioning interaction from a tool to an instrument. The instrument creates awareness through the experience, where the experience of knowing what has been done and felt, as a form of judgment allows the artist to proceed with the extraction and beyond the instrument as a method of doing so. The artifact is not for disposal or utility but part of the process and in the flow of an experience. There does not exist a distinction between doing and undergoing.

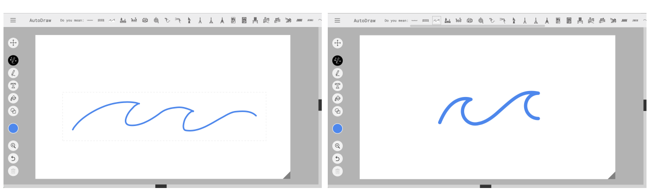
We decipher this integration to not be new but in the likes of other contemporary scholars mention, such as the question of physicality in creativity (Moruzzi 2022) and in transitional terms to Memo Atken's analogy of a "real-time interaction analogous to playing a musical instrument" (Akten, Fiebrink, and Grierson 2019). In fact, extending this analogy, extracting the notion that music now exists, through undergoing, one can decipher an experience such as music without instruments - hearing the sounds of one's surroundings and finding the patterns that arise, to speak of variations (Hui 2021).

Likewise in an analogy cemented in our study, this entails extending AI from a tool to a canvas. For instance, instantiating, FRIDA, a collaborator robotic arm painter that was set to enhance the creativity of the human painter (Schaldenbrand, McCann, and Oh 2022), we find its function for the artistic process does not lend execution to be a conclusion. Circling to Dewey, a technique that can even be better executed by a machine is not "esthetic" and comprehensive to art, yet, in fact, mechanical (Dewey 2005). Hence, if interactions were to embody the esthetic, we must imagine the canvas, as opposed to the tool, and embrace going and undergoing as an embodied experience. In our analogy, the canvas is not the recipient of the robotic arm (tool) but the experience, and beyond the literal analogy, the canvas is the instrument.

Examples of AI Art Instrument

Title: Interpolating Experiences

In the working definition of a tool to assist with ideation and an instrument to aid with expression, we aim to encompass the notion presented above by exemplifying an instantiation of what may emerge if this principle were to be solidified. Hence, we utilize AutoDraw (Motzenbecker and Phillips 2017), an interactive tool that turns sketches into images, to emulate a canvas that mirrors an instrument. We carry henceforth with a wave sketch that is rendered in collaborative completion.



(a) Sketching a doodle (b) Rendering a choice

(c) Pivoting to exploration (d) Synthesized piece

Figure 2: Interpolating Experiences with AutoDraw

To elucidate our conceptual mapping in Fig 2 above, we begin with a wave on an empty canvas as Fig 2a. Following suit, we render the sketch into the options on the top panel of Fig 2b to choose what we intended to draw. We then reiterate the sketch by choosing different meant drawings from the canvas resulting in Fig 2c. Upon diversion and uniformity to intent, colors, and composition, such as sketches that are not turned into meant drawings to present ideas that remain as they are (the splashing of waves to the castle on the left-hand side in Fig 2d), we arrive at our synthesized piece as a conglomeration of the various sketch experiments. Our piece, being enabled by integrated generation, is drawn on the model and lends a progression to AI as a canvas.

Principle 3: Phenomenology of AI

Some scholars have already established instrument-like interaction to occur within a phenomenon known as Embodied Interaction, nearing Principle 2, which as elucidated by Paul Dourish in HCI literature, is the creation, manipulation, and sharing of meaning through engaged interaction with artifacts (Dourish 2001, p. 126). At the core, taking away that experience is what leads to the distinction that one may be artistic and not esthetic, and even to the point that art without esthetic takes away the core ability to utilize experience as judgment and art as expression.

Without judgment that is interwoven with the experience - be it serendipitous or meticulous - the question now becomes how experience may be incorporated within interaction. An experience has pattern and structure, but it is not just doing and undergoing in alteration but consists of them in a relationship. In art, it also incurs an element of freedom in flow (Dewey 2005). However, we acknowledge that this is a far side argument to the notion that 'esthetic' interactions are infinitely intertwined to inherently allow one to form from their experience as judgment. Hence, we further this notion on a comprehensive metric through the phenomenology of Human-Computer Interaction (HCI).

Under the umbrella of experience, a similar alteration is presented to understand interactivity and traces back to the notion of phenomenology. In the steps to a phenomenology of Human-Computer Interaction (HCI) outlined by Daag Svanaes, phenomenology is having an experience that is largely associated with the integration of an outside medium (Svanæs 2000). This alignment opens room for the acknowledgment of such interaction, where it can be extended as not solely an outcome of the experience. That is, embodiment although able, is not comprehensive to its intertwined nature, whereas interaction cites greater avenues, especially for a nascent establishment, to proceed in methods of garnering experience.

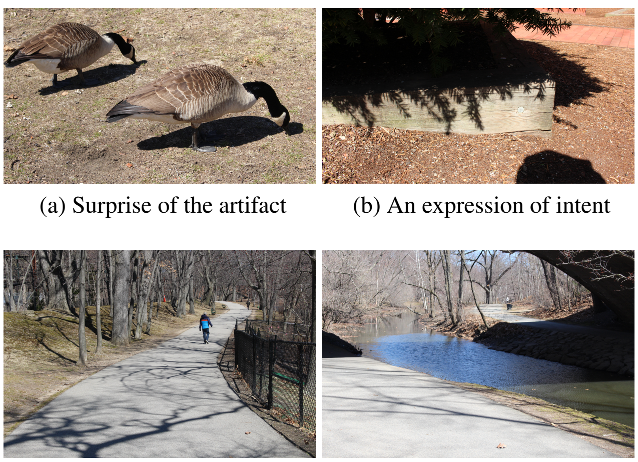
Extending these notions to AI Art, we find the novel aspect of the phenomenology in AI to be its evolution as a medium that encompasses an umbrella of interactions. For instance, as an amalgamation of data, one can curate a dataset and train a model based on the "image" that is fed to its query (Akten 2021), whereas on the other hand, by citing what has already been learned by large models and probing their mechanisms, one can engage in multi-modal interactions such as Prompt Programming to engineer thought mechanisms and express their intent through writing. In these methods lie differences in modality, output, and process that differentiate each interaction, which to its formalization, lends each method to exhibit its own form of experience that establishes its underpinning.

Examples of AI Art Phenomenology

To elicit various experiences, we draw inspiration from shadows as a photographic phenomenon. We choose this medium primarily as images turn reality into a shadow, a memory of what was (Kul-Want 2010, p. 203) and in retrospect, in light of photographic print. In traditional film, photographs were produced by reversing light in the darkroom to cast a shadow onto the photo paper whereby, in this process, every photograph is of a shadow. Shadows also rid bias in the curation of experiences as direct observations.

Title: What AI Art can learn "with" Photography

For this piece, we begin by seeding the experience of photography. Thus, we set out on a quest to capture shadows and embed ourselves in the act of walking around the city in sight of shadows. The session occurred over a 3-mile walk over the course of 4 hours that resulted in a total of 64 shots on a Canon t5i camera. Notable photographs with the associated experiences are presented below:



(a) Surprise of the artifact

(b) An expression of intent

(c) A narrowed view

(d) A tailored eye

Figure 3: Photographic Experiences

To proceed with our process, we then prompted all photographic experiences (Figure 3) in accordance with AI as a "lensless" camera. To build the storyline of our metaphor, we developed our prompt by binding elements that captured metaphorical interest. Devising the prompt as such also takes into account the Meta-Prompt NLP technique that enables the prompt to better inform the model of the experience (Reynolds and McDonell 2021).

Our prompt began with the conditions of a camera. These include location, angles, time of day, subjects, lens type, and aperture. We then utilized vocabulary that recognized the experience as a phenomenological encounter from Minor White, a 20th-century photographer known for his conceptual photographs (Hall and Hoffman 1978). We devise the configurations of our seed seen below:

Seed: 55mm lens at f5.6 taken as an intimate act of contemplative witnessing and co-creation between photographer and subject in a recorded dialogue of light at a given space in time.

To follow suit, we embed each prompt with the learned representation that is inherently symbolic (description of the image) and generate an image with each prompt via Stable Diffusion's Realistic Vision V1.3 model (to emphasize the photographic aesthetic). We present the prompts and corresponding figures as follows:

(a) **Seed** + Two Canada Geese actively feeding on brown grass on a mild sunny afternoon with their shadows underneath as seen from above.

(b) **Seed** + A single occlusion shadow cast on the right of a shrub placed on a raised wooden garden bed surrounded by bark mulch

(c) **Seed** + A meandering narrow asphalt pedestrian path in a park surrounded by leafless trees with sunshine capturing shadows.

(d) **Seed** + Under the bridge of an asphalt pedestrian path next to a calm stream surrounded by leafless trees with sunshine capturing shadows.

Figure 4: Lensless Experiences with associated prompts

Conclusion

In history and trace, photography and AI are drawn together as representations of the objects they capture and depict. In furthering AI Art with photography as its historical predecessor, we recognize and distinguish photographic experiences as carriers of aesthetic representation. We find that this mapping exposes some of the creative affordances of AI Art that might not be available otherwise and do so with principles of interaction to elucidate implementations of such exhibitions. Further implications can also be drawn in extending this viewpoint towards other AI art areas, such as those trained on a corpus of images as the sole artistic input. In our study, we find this approach to be data agnostic, as feeding a "photograph" would not encompass semantic trace, yet leave this as a possibility for inquiry within the instantiated principles and those foregrounded beyond.

Acknowledgements

Many thanks to Pierre V. Tchetgen, Dietmar Offenhuber, and Nathan Miner for the comments and feedback; and Northeastern CSDS for the wisdom in the title.

References

Akten, M.; Fiebrink, R.; and Grierson, M. 2019. Learning to see: you are what you see. In *ACM SIGGRAPH 2019 Art Gallery*, SIGGRAPH '19, 1–6. New York, NY, USA: Association for Computing Machinery.

Akten, M. 2021. *Deep Visual Instruments: Realtime Continuous, Meaningful Human Control over Deep Neural Networks for Creative Expression*. doctoral, Goldsmiths, University of London.

Bodily, P. M., and Ventura, D. 2018. Explainability: An Aesthetic for Aesthetics in Computational Creative Systems. *Proceedings of the 9th International Conference on Computational Creativity, Salamanca, Spain, June 25 – June 29*.

Cetinic, E., and She, J. 2021. Understanding and creating art with ai: Review and outlook.

Dehshibi, M. M., and Adamatzky, A. 2021. Electrical activity of fungi: Spikes detection and complexity analysis. *Biosystems* 203:104373.

Dewey, J. 2005. *Art as Experience*. New York, New York: TarcherPerigee, 1st edition edition.

Divakaran, A.; Sridhar, A.; and Srinivasan, R. 2023. Broadening ai ethics narratives: An indic art view.

Dourish, P. 2001. *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press.

Hall, J. B., and Hoffman, M. E. 1978. Minor White: Rites & Passages.

Hertzmann, A. 2018. Can computers create art?

Hui, Y. 2021. *VARIETIES OF EXPERIENCE OF ART*. University of Minnesota Press. 20–37.

Issak, A., and Varshney, L. R. 2022. Artistic Autonomy in AI Art. *Proceedings of the 13th International Conference on Computational Creativity, Bozen-Bolzano, Italy, June 27 - July 1, 2022*.

Kul-Want, C., ed. 2010. *Philosophers on Art from Kant to the Postmodernists: A Critical Reader*. New York: Columbia University Press, 1st edition edition.

Lotzof, K. 2018. Anna Atkins and the first book of photographs.

Manovich, L., and Arielli, E. 2023. *AI Aesthetics: A Critical Guide to AI in Art, Media and Design*. manovich.net. chapter AI image and Generative Media:.

Mikalonytė, E. S., and Kneer, M. 2022. Can Artificial Intelligence Make Art? arXiv:2104.07598 [cs].

Moholy-Nagy, L., and Molderings, H. 2010. *László Moholy-Nagy: The Photograms: Catalogue Raisonné*. Ostfildern: Hatje Cantz, illustrated edition edition.

Moruzzi, C. 2022. The (Artificial) Physicality of Creativity: How Embodiment Influences Perceptions of Creativity. *Proceedings of the 13th International Conference on Computational Creativity, Bozen-Bolzano, Italy, June 27 - July 1, 2022*.

Motzenbecker, D., and Phillips, K. 2017. AutoDraw by Google Creative Lab - Experiments with Google.

Newton, A., and Dhole, K. 2023. Is AI Art Another Industrial Revolution in the Making? arXiv:2301.05133 [cs].

Reynolds, L., and McDonell, K. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. *arXiv:2102.07350 [cs]*. arXiv: 2102.07350.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schaldenbrand, P.; McCann, J.; and Oh, J. 2022. FRIDA: A Collaborative Robot Painter with a Differentiable, Real2Sim2Real Planning Environment. arXiv:2210.00664 [cs].

Shan, S.; Cryan, J.; Wenger, E.; Zheng, H.; Hanocka, R.; and Zhao, B. Y. 2023. Glaze: Protecting artists from style mimicry by text-to-image models.

Smith, A.; Schroeder, H.; Epstein, Z.; Cook, M.; Colton, S.; and Lippman, A. 2023. Trash to Treasure: Using text-to-image models to inform the design of physical artefacts.

Sontag, S. 2001. *On Photography*. New York: Picador, 1st edition edition.

Sternberger, P. S. 2001. *Between amateur and aesthete: the legitimization of photography as art in America, 1880-1900*. Albuquerque: University of New Mexico Press, first edition edition.

Svanæs, D. 2000. *Understanding Interactivity: Steps to a Phenomenology of Human-computer Interaction*. Norges teknisk-naturvitenskapelige universitet, Institutt for datateknikk og informasjonsvitenskap. Google-Books-ID: WyAzOAAACAAJ.

Ventura, D., and Gates, D. 2018. Ethics as Aesthetic: A Computational Creativity Approach to Ethical Behavior. *Proceedings of the 9th International Conference on Computational Creativity, Salamanca, Spain, June 25 – June 29*.

Zammit, M.; Liapis, A.; and Yannakakis, G. N. 2022. Seeding Diversity into AI Art. *Proceedings of the 13th International Conference on Computational Creativity, Bozen-Bolzano, Italy, June 27 - July 1, 2022*.

I C

ICCC'23

June 19 — 23
Waterloo, Ontario, Canada

14th International Conference
on Computational Creativity

A Association for
Computational
Creativity

IC International
Conference
on Computational
Creativity