Enhancing Human Creativity with Aptly Uncontrollable Generative AI

Iikka Hauhio
Department of Computer Science
University of Helsinki, Finland
iikka.hauhio@helsinki.fi

Abstract

Many generative AI tools produce artefacts, such as text and images, based on the user's instructions given as textual prompts. Users who want to be creative often face a tension: on the one hand, they want the tools to be controllable, i.e., to produce what the user has in mind; on the other hand, they would like the tools to pleasantly surprise them—and in the best case, even enhance their creativity.

In this paper, we address the topics of controllability and creativity in the interaction between a user and a prompt-based generative AI tool, such as a large language model or a text-to-image generator.

We first formalise concepts related to controllability and give measures that allow the description and comparison of tools and use cases. Using the concepts of prompt space and artefact space, of the generative AI tool as a mapping between these spaces, and of a model that the user has of this mapping, we characterise the user's quest for creative artefacts as navigation in the prompt space for inputs that the user thinks are likely to result in artefacts they are aiming at.

We then move on to discuss how the creativity of a user may change based on their interaction with a generative tool. Using our concepts related to controllability, we show how to conceptualise and describe changes that lead to enhancing the user's creativity. This potentially happens when the generative tool is aptly uncontrollable, i.e., it produces artefacts outside the user's original aim, perhaps even something previously unimaginable for the user, and the user updates their subjective ideas about what is possible or desirable.

Introduction

One promise of the generative AI tools is that they allow the users to express themselves in ways they otherwise could not. In order to achieve this goal, they must be controllable, i.e. they must be able to follow the intent of the user, which is not always the case due to their black box nature, unexplainability, and unpredictability (Ganguli et al. 2022; Danilevsky et al. 2020; Hauhio et al. 2023). These qualities are paradoxically both strengths and weaknesses: while they hinder controllability, they also allow the tool to produce works that the user could not have come up with without the tool (Akten 2021).

In this paper, we focus on measuring the controllability and the related qualities in the context of non-interactive tools that take a prompt as input and produce one or more artifacts as output, such as text-to-image models (Rombach et al. 2022). These tools are typically used iteratively: the user adjusts the prompt and generates more artifacts until they decide that the artifacts meet their goals. This process can be described as a search through the artifact space (Choi and DiPaola 2023). In this context, we use the term "controllability" to refer to how easily the user can guide this search to an area of the space that contains the artifacts meeting the goals of the user.

First, we consider controllability and its relation to the mental model of the user. We present a framework for describing the iterative use of tools as a search inspired by the Creative Systems Framework (Wiggins 2019). We propose three *controllability coefficients* as a measure for the controllability of tools. We present an argument that uncontrollability can increase creativity of the interaction between a user and a system. We link uncontrollability to the concept of "serendipity", i.e. happy accidental discoveries. Lastly, we discuss the implications of our work and future research that is required.

Meaningful Human Control

We adapt a definition of controllability proposed by Akten (2021), who suggests loaning the term "meaningful human control" (MHC), previously used in the context of weapons systems, to generative AI discussion. MHC of an AI system is defined through three sufficient and necessary criteria:

1. The system must be able to follow the user's *intent*.
2. The system must be *predictable*.
3. The human must be able to *creatively express* themselves through the system.

This definition can be seen as opposite to Jenning's (2010) *creative autonomy*, which in turn refers to the ability of the system to act "independent of the intentions of its programmer or operator". MHC can also be contrasted with what Akten (2021) calls "button pressing", a term also loaned from weapon system literature: a human that mindlessly accepts the propositions of the computer without questioning them is not meaningfully in control.

Both MHC and button pressing are terms loaned from automatic weapons system literature (Roff and Moyes 2016).

In that context, MHC refers to the "threshold of human control that is considered necessary", with the implication that any lower level of control is unacceptable and unethical. This of course does not apply to the usage of these terms in computational creativity, and we stress that we do not want to bring these connotations with the terms. Controllability is a term that can be applied to multiple kinds of artificial intelligence systems. For automatic weapons systems, it is an ethical requirement. For other kinds of systems, it might not be. Akten's (2021) definition of MHC is intended for analyzing the effects of the system on human creativity, not for determining how ethical or acceptable the system is.

Mental Models

In the heart of controllability is the communication between the user and the tool. First of all, the tool must be able to interpret the user's instructions to be able to follow the user's intent. Predictability is a requirement for this: if the user can predict the actions of the tool, they can also decide how to best communicate the intent to it.

We assume that the user has a mental representation of the tool in their mind that is used to predict its behavior. How this model develops is crucial to the perception of the user of the tool. Wardrip-Fruin (2007) proposes three alternatives for how the model might evolve, named after three interactive systems: the ELIZA effect, the TALE-SPIN effect, and the SimCity effect. The two former effects have negative consequences for the user's perception of the system, while the last effect can be seen as more positive and desirable. The ELIZA effect refers to a situation in which the users first perceive the system as too intelligent, i.e. their mental model is inaccurate by overestimating the complexity of the system. The TALE-SPIN effect is the opposite of this: for some reason, the users underestimate the complexity of the system and judge its output random or non-meaningful, and are thus unable to appreciate its processes. The SimCity effect, on the other hand, describes a situation in which the users gradually learn a more and more accurate mental model of the system by communicating with it.

In addition to the user's mental model changing, the system might also change. For example, the SimCity game used as an example by Wardrip-Fruin (2007) begins with simplified game mechanics, and the more complex mechanics are progressively enabled as the player learns to play. In the context of computational creativity, a similar idea has been previously presented by Cassion, Ackerman, and Jordanous (2021) who introduce the concept of *humble creative machines* that gradually adjust their behavior to "meet the user at the level of their expertice".

Since the user's mental model, and in some case the tool as well change over time, the system's predictability and controllability also change. It is thus not possible to determine *the* controllability of the system objectively. However, controllability can be measured for a specific user on a specific time point. In this paper, we generally assume that the user's mental model or the system do not change significantly during measurements. It is important to keep in mind that multiple measurements are required to get a more complete understanding of the system.

Creativity as a Search in Space

The iterative use of generative artificial intelligence tools can be intuitively modeled as a search in the artifact space (Choi and DiPaola 2023). To formalize this behavior, we define a framework based on the Creative Systems Framework (CSF) (Wiggins 2006; 2019) extended to include both the user and the system. While the CSF originally models the search as performed by the system, in our framework the search is performed by the user who uses the system as a tool.

The CSF defines creative acts as iterative processes that search the conceptual space based on a set of rules and evaluations (Wiggins 2006; 2019). More specifically, Wiggins defines a universe $\mathscr{U}$, which is a space containing every concept. During the creative process, a subset of the universe defined by rules $\mathscr{R}$ is explored following traversal rules $\mathscr{T}$ and evaluation rules $\mathscr{E}$. The exploration is an iterative process in which an interpreter function $\langle\!\langle \mathscr{R}, \mathscr{T}, \mathscr{E} \rangle\!\rangle(\cdot)$ is used to map the previous concept into the next concept, thus finding new concepts in the space.

We extend the CSF by defining two conceptual spaces inside the universe: the *prompt space* $\mathscr{P} \subset \mathscr{U}$ and the *artifact space* $\mathscr{A} \subset \mathscr{U}$. The generative AI tool G is a mapping $G : \mathscr{P} \to \mathscr{A}$.[1] We also define $\mathcal{M} \sim G$ as the mental model of the user.

While the user uses the tool, they explore the prompt space $\mathscr{P}$ by using some internal traversal strategy $\mathscr{T}_{\mathcal{M}}$. However, as their objective is not to create prompts but to create artifacts, they will therefore use artifact evaluation rules $\mathscr{E}_{\mathscr{A}}$ to guide their search. Their goal, i.e. the space of concepts they are looking for, is defined by the rules $\mathscr{R}_{\mathscr{A}}$. The interpreter function they use is thus $\langle\!\langle \mathscr{R}_{\mathscr{A}}, \mathscr{T}_{\mathcal{M}}, \mathscr{E}_{\mathscr{A}} \rangle\!\rangle(\cdot)$.

$\mathscr{T}_{\mathcal{M}}$ is a traversal strategy that relies on the mental model. $\mathcal{M}$ is like a lookahead function that allows making educated guesses about possible new prompts, thus making it possible to efficiently search the space. As the user continues to use the tool, $\mathcal{M}$ will change as the user gains more skill and knowledge of how the tool works.

Generally, the better the model $\mathcal{M}$ is, the better the user is at controlling the tool, since they can more easily make transformations in the artifact space by changing the prompt. Thus, the controllability of the tool is dependent on the accuracy of $\mathcal{M}$. However, not all programs are controllable even if the model is fully accurate due to non-determinism, side effects, and other aspects that are outside the influence of the user. Moreover, some programs are just incapable of generating the artifacts the user wants, thus making it impossible to control the program.

The artifact space can be divided into three possibly overlapping areas (the sets have been named by us, but the notation after the equals sign comes from (Wiggins 2006)):

[1] This is not completely accurate as typically a generative AI tool is not a function but a distribution. However, this distribution is very often sampled using a pseudo-random sampling algorithm with a known seed, making the tool a function. For notational simplicity, we have chosen to simply represent the tool as a function from the prompt space to the artifact space.

1. $C = [\![\mathscr{R}_{\mathscr{A}}]\!](\mathscr{A})$ (conceptual set) is the set of artifacts that fulfill the goal of the user as specified by $\mathscr{R}_{\mathscr{A}}$.
2. $V = [\![\mathscr{E}_{\mathscr{A}}]\!](\mathscr{A})$ (valued set) is the set of artifacts that are valuable as specified by $\mathscr{E}_{\mathscr{A}}$.
3. $T = \langle\!\langle \mathscr{R}_{\mathscr{A}}, \mathscr{T}_{\mathcal{M}}, \mathscr{E}_{\mathscr{A}} \rangle\!\rangle^{\diamond}(\{\bot\})$ (traversable set) is the set of artifacts the user can actually reach by using the strategy $\mathscr{T}_{\mathcal{M}}$.

During the search, the user will try to find artifacts that belong to both C and V, i.e. artifacts that both satisfy their goal and are valuable. It should be noted that, like the model $\mathcal{M}$, these sets might also change over time as the user's perception of what is the goal and what is valuable changes. In the following sections, we assume that they do not change significantly during a single session. If the sets are being measured experimentally, the goal of the user should be specified beforehand to prevent it changing.

Measuring Controllability

Meaningful Human Control requires the user to be able to (1) predictably produce artifacts that (2) follow their intent and (3) through which the user can creatively express themself (Akten 2021). In terms of the framework described in the previous section, the condition (2) is fulfilled if the generated artifacts belong to the conceptual set C. (3) further requires that the artifacts must be good enough to be able to used for the user's self-expression, thus the artifacts should also belong to V. Finally, the condition (1) is fulfilled if the generated artifacts predictably belong to these sets, which we will measure as the number of artifacts belonging to the sets divided by the total number of generated artifacts.

Inspired by the concept of *curation coefficient* as defined by Colton (2012), we define three coefficients for formalizing the above idea: the *meaningful control coefficient*, the *control coefficient*, and the *value coefficient*. The formal definitions are given below.

Definition 1. Let $A \subset T$ be the set of artifacts generated during a session. The following coefficients are defined for this session:

$$\text{MeaC} = \frac{|A \cap C \cap V|}{|A|} \tag{1}$$

$$\text{ConC} = \frac{|A \cap C|}{|A|} \tag{2}$$

$$\text{ValC} = \frac{|A \cap V|}{|A|} \tag{3}$$

The *control coefficient* refers to the share of artifacts containing the concept desired by the user. If the coefficient is high, the user is able to communicate their concept to the system, the system is able to understand it, and so the system is controllable in some sense, although it does not fulfill the requirements of meaningful control yet. The *value coefficient*, on the other hand, refers to the share of artifacts that are valuable, but not necessarily containing the concept specified by the user. Finally, the *meaningful control coefficient* refers to the share of artifacts belonging to the intersection of the two previous sets. If the last coefficient is high, all three conditions of MHC are satisfied.

In addition to the three coefficients we defined, Colton's (2012) curation coefficient can be defined as the number of selected artifacts. By "selected", we refer to artifacts that the user actually intends to use after the session. For example, if the user generates multiple nearly identical artifacts and chooses to use only one of them, the number of selected artifacts is one, even though any of the artifacts could have been selected.

Definition 2. Let S be the set of artifacts selected by the user for further use. Then

$$\text{CurC} = \frac{|A \cap S|}{|A|} \tag{4}$$

In this paper, we assume that the systems being evaluated are not interactive, that is, they work by giving one or more outputs for a given input. While artifacts can be generated iteratively by changing the input and generating new artifacts, the system is not interactive in the sense that it would ask the user anything after the initial input is given. This corresponds to how AI image generators like Midjorney and Stable Diffusion (Rombach et al. 2022) work when generating images from scratch, but not to how some more interactive features like image inpainting work, or to how explicitly interactive systems like the ChatGPT (OpenAI 2023) work. It is also not possible to calculate the coefficient for regular text and image editors, although it can be applied to most filters and effects an image editor has.

In addition to non-interactivity, we assume that each use of the system will have a *goal*. Together, all runs of the program that have the same goal are called a *session*. This is assumption is necessary for us to be able to determine the number of artifacts generated: if the user is just playing with the system and happens to come across a good artifact, we cannot determine which of the previous artifacts belonged to the same session since the goal was not predetermined. While this assumption does not hold generally, we can enforce it when measuring the coefficients empirically.

Thus, measuring the coefficients of a non-interactive system works as follows: before anything is generated, the user must state the goal of the current session. After that, they will proceed to generate artifacts, possibly in multiple runs iteratively. At the end of the session, the user will choose which artifacts are valuable and satisfied their initial goals and which did not. The meaningful control coefficient is the number of valuable and goal-fulfilling artifacts divided by the total number of artifacts.

When reporting an empirically measured coefficients, the set of goals used and the skill of the user should be clearly indicated, as they might affect the coefficients radically. In addition to that, the method for measuring value of artifacts and the threshold that is used for considering an artifact "valued" should be specified.

For the concept set, Wiggins (2006) specifies a constant threshold of 0.5, i.e. C contains the artifacts for which $[\![\mathscr{R}_{\mathscr{A}}]\!](a) > 0.5$. However, in some situations it might be meaningful to instead define multiple thresholds, and thus

multiple concept sets. Similarly, one might define multiple valued sets, and multiple control and value coefficients. I.e. instead of reporting a single value coefficient, one could report a "high value coefficient" with a high threshold of value, a "moderate value coefficient", a "low value coefficient", and so on.

Uncontrollability and Creativity

In so far, we have assumed that the goals of the user are static and unchanging, i.e. that the user has a concept for the artifact already in their mind, and that the role of the generative AI tool is merely to help find or implement a valuable instance of that concept. If this is the case, uncontrollability and unpredictability can be viewed as hindrances preventing the user from finding that instance they are looking for.

However, if it happens that during the search the system generates something outside the set C, i.e. something not fulfilling the goal of the user, and that something is valuable, the user might reconsider and change their goals ($\mathscr{R}$). This can be seen as transformative creativity, as the whole concept set has changed (Wiggins 2019, section 2.4.9). Furthermore, as the user learns to use the tool to generate these new kinds of artifacts by figuring out what kinds of inputs produce them, they also change their mental model $\mathcal{M}$ and by that their traversal strategy $\mathscr{T}_{\mathcal{M}}$, also a form of transformational creativity.

In addition to transformativity, we also argue that uncontrollable systems also simply produce more novel artifacts (at least to the user), since their output will be something the user did not think about beforehand, or even something totally unimaginable to the user in the most extreme cases.

Thus, the model's uncontrollability and unpredictability, whether caused by their stochasticity, chaoticity, or the imperfect mental model of the user, is not necessarily a hindrance, but a source of increased creativity. However, it is clear that the system cannot be *completely* uncontrollable: if the user cannot influence the system at all, their own personal goals and their transformations do not play a part in the creative process anymore. Therefore, the generative AI tools need to be *aptly uncontrollable*, i.e. allow some guidance from the user while still be able to sometimes force the user to go to unmapped territories of the conceptual space.

Another viewpoint to the controllability of the system is to view it as an instance of the exploration–exploitation dilemma. A tool that is completely unpredictable only allows *exploring* the conceptual space without giving the user an opportunity to decide which areas of it the tool should focus on, i.e. *exploit*. On the other hand, if the tool is completely controllable, it is up to the user to perform the exploration.

Uncontrollability and Serendipity

The idea of uncontrollability being beneficial for creativity is linked to the concept of *serendipity*, i.e. fortunate accidental discoveries that produce valuable outcomes. Pease et al. (2013) list three dimensions of serendipity: 1) chance (discovery is accidental), 2) sagacity (skill of the discoverer), and 3) value (of the result). We argue that these three dimensions are present in aptly uncontrollable systems: 1) the element of chance comes from the uncontrollability resulting in unpredictable interactions between the user and the system; 2) the element of sagacity comes from the ability of the user to recognize promising artifacts, change their mental model, and guide the search; and 3) value comes from the ability of the system to produce valuable artifacts. Thus, the system, the user, and their interaction are all required for the serendipitous process to occur. This highlights the co-creative nature of using aptly uncontrollable systems.

The term serendipity is often used in recommender system literature (Kotkov, Wang, and Veijalainen 2016). A recommender system should not just give suggestions based on the previous interests of the user, but also suggestions different to those the user has previously reviewed. There are many similarities between recommender systems and generative AI tools: both of them ultimately produce useful artifacts for the user, one by searching a database and another by generating them. A key difference between the two in the context of our paper is that we have mostly assumed that the system is unchanging, while recommender systems usually have a model of the user that they update during use to increase the accuracy of their predictions of user interests. We believe that many recommender system algorithms might be applied to generative AI tools, but leave further analysis to future research.

Discussion and Conclusions

In this paper, we presented a framework for describing the iterative use of generative AI tools inspired by the Creative Systems Frameworks (Wiggins 2019). We then defined three coefficients, the *meaningful control coefficient*, the *control coefficient*, and the *value coefficient*. Lastly, we analyzed the implications of uncontrollability to the creativity of the system and concluded that a correct balance of controllability and uncontrollability can increase the transformative creativity of the process.

We believe that measuring the coefficients of existing systems would be beneficial, even though they are highly situation-dependent. The coefficients would enable us to compare the controllability of different systems and configurations, which would be beneficial to the users of the systems, their developers, and researchers. To make different test results comparable with each other, standardized measurement methods and test environments should be devised.

The ideal values of the coefficients depend on the use case of the system. Is the purpose is to maximize human creativity and self-expression, the meaningful control coefficient should be as high as possible. If, on the other hand, the purpose is to produce novel and valuable artifacts, the control coefficient might be lower, as it forces the user to encounter new kinds of novel artifacts during the search.

In all cases, it is desirable to have a high value coefficient. One might argue that, in the context of creativity, it is more important that the system produces valuable output than that the system follows the instructions given to it, and thus when training generative AI models, more focus should be placed on their ability to assess the value of their output, and less on their ability to follow the prompt. However, we

note that the in some cases it might be enough for the system to be aptly uncontrollable without actually producing valuable end results. For example, the user might use the system to gain novel ideas of concepts, and then use another system to implement those ideas. In this case, the artifact evaluation rules $\mathscr{E}_{\mathscr{A}}$ of the user correspond to the value of the concept instead of the value of the artifact itself.

Even though we defined formalized the ideas presented in this paper only for non-interactive systems, we note that they do apply to other kinds of systems as well. We leave the analysis of these systems for future research.

Even though we have argued for the benefits of uncontrollability, we stress that in general, controllability of generative AI tools is important. The number of iterations it takes for a user to find a valuable artifact satisfying their goals correlates directly with the monetary and environmental costs of generating artifacts (Utz and DiPaola 2023). We believe that uncontrollability, while beneficial in some situations, should be limited to the context in which the benefits outweigh the costs. In particular, if uncontrollability correlates with lessened value, it is not likely to have an effect on the user's goals and is thus unnecessary.

Acknowledgments

IH is funded by the doctoral programme in Computer Science at the University of Helsinki.

References

Akten, M. 2021. *Deep visual instruments: realtime continuous, meaningful human control over deep neural networks for creative expression*. Ph.D. Dissertation, Goldsmiths, University of London.

Cassion, C.; Ackerman, M.; and Jordanous, A. 2021. The humble creative machine. In *Proceedings of the 12th International Conference on Computational Creativity*.

Choi, S. K., and DiPaola, S. 2023. Art creation as an emergent multimodal journey in artificial intelligence latent space. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, 247–253.

Colton, S.; Wiggins, G. A.; et al. 2012. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence*, volume 12, 21–26.

Danilevsky, M.; Qian, K.; Aharonov, R.; Katsis, Y.; Kawas, B.; and Sen, P. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 447–459.

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.

Hauhio, I.; Kantosalo, A.; Linkola, S.; and Toivonen, H. 2023. The Spectrum of Unpredictability and its Relation to Creative Autonomy. In *Proceedings of the 14th International Conference on Computational Creativity*.

Jennings, K. E. 2010. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines* 20(4):489–501.

Kotkov, D.; Wang, S.; and Veijalainen, J. 2016. A survey of serendipity in recommender systems. *Knowledge-Based Systems* 111:180–192.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Pease, A.; Colton, S.; Ramezani, R.; Charnley, J. W.; and Reed, K. 2013. A discussion on serendipity in creative systems. In *Proceedings of the Fourth International Conference on Computational Creativity*, 64–71.

Roff, H. M., and Moyes, R. 2016. Meaningful human control, artificial intelligence and autonomous weapons. Technical report, Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Au-Tonomous Weapons Systems, UN Convention on Certain Conventional Weapons.

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.

Utz, V., and DiPaola, S. 2023. Climate implications of diffusion-based generative visual ai systems and their mass adoption. In *Proceedings of the 14th International Conference on Computational Creativity*.

Wardrip-Fruin, N. 2007. Three play effects–eliza, tale-spin, and sim city. *Digital Humanities* 1–2.

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

Wiggins, G. A. 2019. A framework for description, analysis and comparison of creative systems. In *Computational Creativity*. Springer. 21–47.