# *Who's driving? Exposing human self-deception in creative human-AI collaboration*

**Špela Vintar**[1,2] and **Mojca Brglez**[1,2] and **Damjan Popič**[1] and **Jan Jona Javoršek**[2]
[1] Faculty of Arts, University of Ljubljana
Aškerčeva 2, Ljubljana, Slovenia
{damjan.popic}@ff.uni-lj.si
[2] Jožef Stefan Institute
Jamova 39, Ljubljana, Slovenia
{spela.vintar,mojca.brglez,jona.javorsek}@ijs.si

**Abstract**

With the surge of AI tools designed to help with various creative tasks from writing, video production and music composition to crafts and performing arts, collaboration with AI has become standard practice for many creative professionals. We present a study of ideation and authorship on a dataset of 125 structured interviews with creative professionals describing their use of AI. We hypothesize that the majority of creative professionals underestimate AI's contribution to their ideation processes and the authorship of the final creative product. We seek to prove our hypothesis through three independent methods: Firstly, we perform a corpus-based analysis of the interviews by extracting and manually categorizing verb and noun phrases indicating human vs. AI agency. By assigning scores to individual parts of each interview, we identify the dissonance between self-perceived and actual AI roles. Secondly, we apply a word-embeddings-based approach, including clustering and direct distance-based similarity metrics to classify the phrases into either human or AI-driven creation. Finally, we collect human judgements by asking human experts from similar creative professions to assess AI's role in ideation and authorship for each interview. Results indicate that humans often underestimate AI's contribution, but also that defining the share of AI-agency in human-AI collaborative works is sometimes indeed ambiguous.

## Introduction

In December 2025, Anthropic released a study on how professionals from different fields use AI (Handa et al. 2025). The study was based on 1,250 interviews conducted between an AI agent and people who regularly use AI in their work, and the sample involved—aside from 1,000 general workforce and 125 scientific participants—125 creative professionals engaging in a broad spectrum of activities ranging from music composition, screenwriting and fiction writing to design, photography and video production. The dataset was made public on the HuggingFace platform[1], and Anthropic's analysis of the interviews by subgroups presents findings related to the workflows, usage patterns, attitudes, and concerns of the informants.

[1] https://huggingface.co/datasets/Anthropic/AnthropicInterviewer

As researchers with an interest in computational creativity and human-AI collaborative creation (HAIC), we started reading the interviews with creative professionals to learn more about their collaboration paths and co-creation processes. Within this hugely varied set of first-hand accounts, we observed an interesting phenomenon which grew into the main motivation for this study: Guided by the AI interviewer, participants would first describe their line of creative work, list the ways in which they use AI and illustrate the collaboration with a detailed account of a recent project. In some cases, these accounts delivered the impression that a large portion of the creative work, both from the conceptualization and ideation stage as well as the actual execution, was outsourced to the machine. However, when asked about *the driver of the creative process*, the prevailing sentiment was one of *undisputed human control*, with statements such as "I am fully in charge," "I make all the decisions," "I drive the creative process." We believe such discrepancies between the extensive involvement of AI models and the users' assertions of full control do not stem from intentional misrepresentation but suggest a form of *self-deception* (Gur and Sackeim 1979; Mele 2001b), driven by their desire to remain in control, as "[f]or the most part, people who deceive themselves into believing that *p* want it to be the case that *p*" (Mele 2001a, p. 96).

While acknowledging the complexity of the authorship concept in HAIC (see the following section), we were interested in whether a potential mismatch between the user's self-reported AI-use on the one hand and their impressions about control and creative autonomy could somehow be revealed. With the interviews as our only source of information and no access to the anonymous participants or their works, we present three experimental methods aimed at assessing the role of AI in these HAIC use cases, and comparing it to the users' "control" claims. The first method employs traditional corpus-based techniques and relies on manually annotated examples, the second is based on neural Natural Language Processing (NLP) methods using topic modeling and word embeddings, and the third uses human experts, more specifically creative professionals from related fields, who contribute ideation and authorship judgments on each individual use case.

The remainder of this paper is structured as follows: We

give a very brief summary of related work focusing on human-AI co-creativity and discussions of authorship, but also presenting psychological aspects of authorship attribution in human-human collaboration. We proceed with a presentation of the dataset and the pre-processing steps performed before computational analysis. The next three sections present our approaches to identifying egocentric bias or instances of AI users' self-inflation, whereby the methods and results are presented separately for each section, then compared in the Discussion section.

## Related work

The nuances of human-AI collaborative creation (HAIC) have been debated by various authors and have assumed philosophical, psychological, legal, cognitive, or interdisciplinary perspectives. The first observable stance is that while computers can be creative and can play a significant role in HAIC, they can rarely be considered (co-)authors. The foundational work of Boden in her revised edition of *The Creative Mind* justifies this position through the lack of human-like motivation and intention (Boden 1992). A much more recent contribution to the theory of co-creativity and authorship is by Liu and Huang (2025) who introduce the concepts of artistic imagery thinking and control, both of which are central elements in authorship attribution. While the distribution of control in HAIC scenarios may vary, the authors claim that AI does not engage in artistic imagery thinking, and thus cannot be attributed authorship. A similar view from a strictly legal perspective is presented by Ginsburg and Budiardjo (2018), who analyze various jurisdictions and conclude that the purpose of copyright is the protection of human creative production, thus a creative product is copyrightable with sufficient human involvement, otherwise authorless.

An opposite stance advocating a shift in legal practices regarding AI is presented by Abbott and Rothman (2022), whose position is that attributing authorship to AI would increase transparency and even protect human authors. A similarly inclusive view of co-creation can be seen in Boo et al.'s (2025) comprehensive overview of HAIC theories, who present a multifaceted framework for collaborative creative processes. Despite the large body of work debating human-AI co-creation (see also Kantosalo and Toivonen 2016; Bown et al. 2020; Nikrang and Kiesenhofer 2025; Condorelli and Berti 2025; Karimi et al. 2018) and the positions briefly presented above, we argue that the issue of authorship remains complex and dynamic, especially in view of the rapid advances in GenAI for text, image, video and music generation. For the research question we explore, namely a possible dissonance between self-perceived and objectively attributable authorship, studies of human-human co-creation seem particularly relevant.

In behavioural and psychological studies, the phenomenon of overestimating one's own contribution in a collaborative setting can have different motives: egocentric bias makes us exaggerate our inputs because our own work is more familiar and present to us (Ross and Sicoly 1982); other people's creativity may instill envy and ostracism (Breidenthal et al. 2020); or misattribution may occur due the so-called Matthew/Matilda effect, referring to status- or gender-based discrimination often observed in science (Rossiter 1993). Perhaps the AI-coauthor in some cases resembles a ghost writer (Pruschak and Hopp 2022), an invisible and uncredited provider of ideas, feedback and generated content. Although such cases may not violate existing legal standards of authorship, they raise parallel ethical concerns regarding credit allocation and transparency.

## Dataset and Pre-processing

The entire Anthropic dataset consists of 1,250 semi-structured interviews with various types of professionals in which they discuss their use of AI with an AI interviewer. From this dataset, we only analyze the 125 interviews with professionals from creative fields, such as designers or writers[2]. The length of the interviews varies from 1,000 to 2,500 words, and they are anonymized for person, project or company names which could reveal the identity of the participant.

Each interview comes from a semi-structured dialogue consisting of questions that recur across all of them but may be adjusted on the spot to maintain a natural conversational flow. We segment the transcripts into sections[3] and restrict our analysis to the four sections most relevant to the process of creation and HAIC. The analyzed sections are:

- `basic job description`: In this section, the participants respond to the AI agent's question "Could you tell me a bit about your creative work and what a typical project looks like for you?"
- `walkthrough`: Here, the participants describe their ways of using AI by responding to: "Walk me through how AI fits into your creative process."
- `project example`: The agent asks the participant to describe a specific recent project where AI was used, and participants provide details about the role AI played.
- `dynamic`: In this section, the participants respond to the question: "Who or what is driving the creative decisions in this process?" and describe the dynamic of the collaboration.

All analyses in this paper are computed on interviewee turns only.

### Pre-processing and annotation

All transcripts were processed with the Stanza pipeline[4] to obtain linguistic annotations used for slicing and downstream aggregation:

[2]The dataset is already annotated with the profession type, distinguishing between creatives (N=125), scientists (N=125), and the general workforce (N=1000).

[3]We identified a total of 8 recurring sections apart from the short introductory consent to the study: basic job description, a brief walkthrough of the creative process, a recent project example, the dynamics of human–AI collaboration, the most changed aspects in the field due to AI, potential concerns about AI, future prospects, and additional comments.

[4] `https://stanfordnlp.github.io/stanza/pipeline.html`

- token content and normalization: `text`, `lemma`
- morphosyntax: `upos`
- dependency structure: `head`, `deprel`

### Creative type classification

To enable comparisons across professional domains, each interview is assigned a creative-type label.[5] The final typology distinguishes nine types: Arts & crafts, Design, Film & video, Music, Photo, Screenwriting, Web content creation, Writing fiction, Writing miscellaneous (refered to as `arts_crafts`, `design`, `film_video`, `music`, `photo`, `screenwriting`, `web_content`, `writing_fiction`, `writing_misc` in figures).

## Corpus-based Analysis

### Rationale

Our aim is to quantify how interviewees allocate agency between humans and AI, and how this allocation varies across interview sections and creative types. The scoring is corpus-based in the strict sense that it is computed from dependency-parsed text and can be traced back to recurring lexico-grammatical patterns rather than inferred from sentiment or topics (Kilgarriff 2001; Dunning 1993; Hardie 2014). The design goal is therefore twofold: the measures are comparable across interviews, but also auditable—for any aggregate difference we are able to point to the concrete triplets that drive it.

### Units of analysis

Single words are too ambiguous to code for agency. We therefore operate on dependency-derived lemma-level triplets. For each verbal predicate ($\texttt{UPOS} \in \{\texttt{VERB}, \texttt{AUX}\}$), we extract an agent/subject slot and, where available, a patient/object slot. In addition, we optionally record an embedded action (verbal `xcomp`/`ccomp`) when it is itself verbal. Triplets are stored on the lemma level to reduce sparsity and to support aggregation across interviews, while short phrase spans are retained for inspection. From the extracted triplets, we analyze the unique triplets that appear in the dataset at least 2 times.

### Labeling triplets

Each unique triplet type is assigned a binary label based on whether it conveys human or AI agency, whereby the use of AI is considered AI agency even if the grammatical subject is human (e.g., *I use [the] model*):

- `HUMAN`: the agent is human-framed
- `AI`: the agent is AI-framed

When a triplet is impersonal or unclear (e.g., placeholder subjects), we do not force a binary assignment. Such cases are treated as unlabeled and excluded from labeled-share scoring. Because this conservative policy can affect the amount of material it labels across sections or creative types, we monitored the proportion of labeled vs. unlabeled triplets to ensure that the scoring relies on a substantial portion of the data. A small set of illustrative triplets is shown in Table 1.

| Triplet | Label |
|---|---|
| `I :: use :: AI` | `AI` |
| `I :: use :: model` | `AI` |
| `I :: have :: idea` | `HUMAN` |
| `I :: make :: decision` | `HUMAN` |

Table 1: Illustrative examples of dependency-derived triplets and their manual HUMAN/AI labels.

| Section | Representative triplets |
|---|---|
| *basic job description* | `I :: use :: AI` |
| *dynamic* | `I :: drive :: decision;` |
| | `I :: make :: decision;` |
| | `I :: have :: idea` |

Table 2: Contrasting section-salient triplets consulted during codebook refinement.

In this scheme, actions that delegate work to an AI system (e.g., `I :: use :: AI`) are labeled as `AI` agency even when the grammatical subject is human, because the operative step of the process is attributed to the system rather than the user. We operationalize this distinction through manual triplet labeling, which yields a "codebook" used for two purposes: first, to compute AI-agency scores directly, and second, to serve as a basis for the concept vectors described in the next section.

### Using triplet lists to refine the codebook

To complement overall frequency lists, we also inspected triplets that were especially characteristic of particular interview sections (Table 2). In practice, the strength of such section-specific patterns was uneven: *dynamic* yielded the clearest distinctive triplets, whereas *basic job description* was comparatively less distinctive. This contrast was useful both during codebook refinement, where it helped identify constructions tied to specific discourse modes, and later in interpreting section-based score differences, since some sectional patterns are driven by strongly section-specific constructions while others are supported by more diffuse evidence.

This diagnostic layer was thus not used as a separate scoring procedure, but it helped relate aggregate section-level differences back to the constructions that produced them.

### Computing the agency score

The AI-agency score is defined as the proportion of labeled triplet occurrences in which agency is attributed to AI rather than to humans. In other words, for any subset of the corpus

[5] Anthropic's analysis of the interviews distinguished six types of creatives, including writers, visual artists, craftspeople, designers, filmmakers, and game developers, but these labels and splits are not publicly available. After inspecting the dataset instances, we opted for a finer-grained distinction among creative professions. We split and labeled the creative types using clustering techniques combined with manual inspection. In this step, we have also eliminated 5 interviews where the interviewee was not in fact performing creative work or belonged to a more niche profession (profesional voice actor, board game designer).

$X$, the score is the number of AI-labeled triplet occurrences divided by the total number of labeled triplet occurrences in that subset. Let $N_A(X)$ and $N_H(X)$ denote the counts of AI and HUMAN triplet occurrences in $X$. The AI agency share is:

$$\text{AIShare}(X) = \frac{N_A(X)}{N_A(X) + N_H(X)}. \quad (1)$$

In the analyses reported here, this score is used primarily for section-level and creative-type-level aggregates, as illustrated in Figure 1.

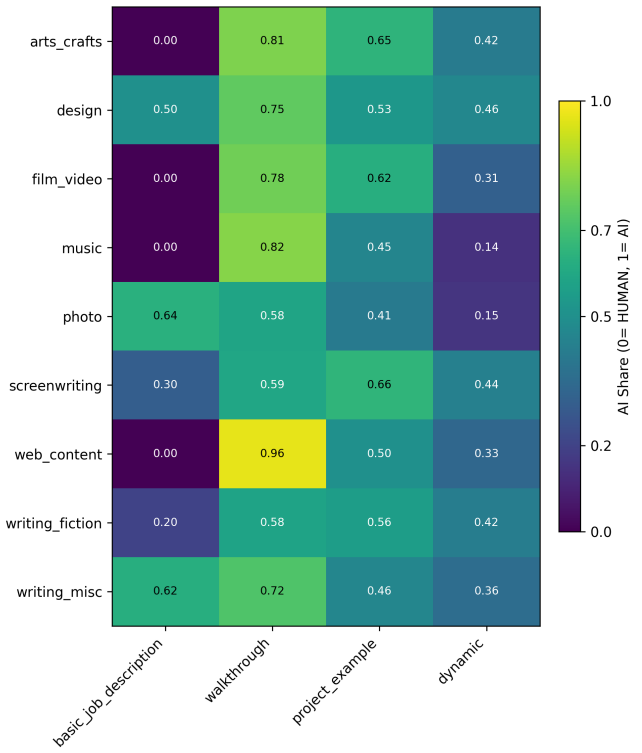


Figure 1: AI-agency share by interview section and creative type. (Higher value and lighter = higher AI share. Human agency prevails in sections providing the basic_job_description, while AI agency is visibly dominant in the walkthrough sections, most notably for web_content creatives.)

### Aggregation by transcript section and creative type

To compare the AI-shares in different transcript sections and across professional domains, we aggregate scores for these subgroups. Figure 1 reveals the main trend, namely that human agency prevails in the initial *job description*, then we see a strong prevalence of AI-agency in the *walkthrough* section, a mixed human-AI picture in the *project example* section, and again a visible return to human agency in the section where users describe the collaboration *dynamic*.

To avoid unstable estimates, we report agency shares only for aggregates with sufficient labeled evidence. In the comparison shown in Figure 1, triplet occurrences are pooled by creative type and interview section, and the plotted values represent the share of AI-labeled triplets within each creative-type-by-section subset. This aggregation provides a more robust basis for comparison than interview-level scoring, which may be sparse for individual cases. Throughout, we retain the underlying triplet lists so that the reported patterns remain transparent and traceable to their underlying linguistic realizations.

The approach is dependency-based, meaning that grammatical relations define structured slots (agent, patient, embedded action) over which recurring lexico-grammatical patterns can be aggregated, broadly analogous to dependency-based slot-filler approaches such as word sketches, i.e., automatic corpus-derived summaries of a word's grammatical and collocational behaviour (Kilgarriff and Tugwell 2001).

## Clustering and Semantic Distance Measures

In this section, we describe our distributional semantic approaches for automatically assessing the contribution of AI to the creative process. We retain the analysis at the triplet level for two reasons: (1) to preserve comparability with the corpus-based approach, and (2) to avoid fragmentation of the embedding vector space due to the specificities of creative professionals' jobs, projects, and applications. This section presents two main methods: coarse clustering based on the most frequent triplets, and a finer distance-based similarity approach applied to all triplets in the analyzed sections.

### Clustering

To obtain an initial coarse-grained overview of the main creativity-related themes in the data, and to estimate the relative contributions of users and AI, we first apply clustering of the data. In this step, we only use unique triplets that include the most frequent verbs in the dataset [6].

Triplets are encoded using the sentence transformer model all-MiniLM-L6-v2[7]. First, we apply the $k$-means algorithm combined with PCA dimensionality reduction to 100 dimensions, and determine the optimal number of clusters $k$ using silhouette analysis and visual inspection of the cluster projections. Both methods converge on $k = 3$. Figure 2 shows the resulting clusters and their representative triplets. Based on the dominant triplet patterns, we can interpret the blue and green clusters as reflecting AI use and HUMAN agency, respectively.

We also used BERTopic (Grootendorst 2022) to perform HDBSCAN density-based clustering to identify semantically coherent groupings, without requiring a pre-specified number of clusters. Again, we applied it to all triplet instances of the top 200 verbs, which yielded two main clusters alongside a set of outliers. We interpreted the two clusters similarly to the two $k$-means clusters, namely, one

[6]This triplet set slightly differs from the one described in the corpus-based analysis section. Because we are not constrained by manual labeling, we apply a frequency threshold only to the predicate (rather than the entire triplet) and do not lemmatize the data in order to preserve richer semantic representations.

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

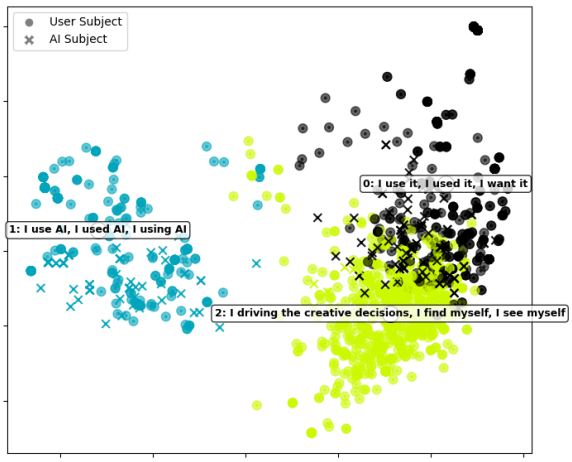


Figure 2: $K$-means clusters, PCA-reduced dimensionality. (Axes represent the first two dimensions. Blue cluster on the left (1): `AI` use, Bright green cluster on the bottom right (2): `HUMAN` agency.

seems to correspond more to `HUMAN` agency and the other to `AI` use.

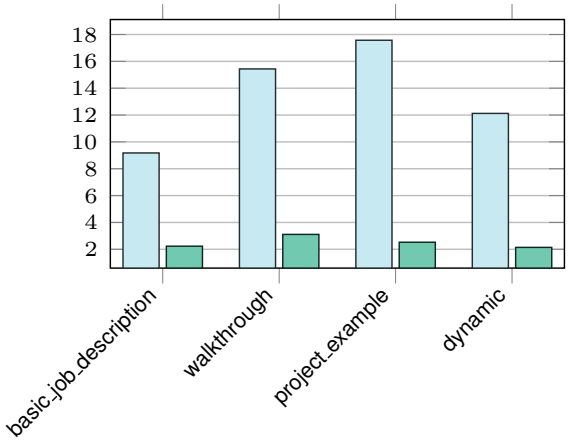


Figure 3: Average BERTopic distributions per section. (Brighter blue = `HUMAN`, green = `AI`. The vast majority of the triplets are placed in the `HUMAN` topic.)

Figure 3 shows the average number of triplets per section assigned to the topic clusters identified by the fitted BERTopic model. The large majority of the triplets are placed in the first topic, meaning `HUMAN` agency triplets seem to dominate all of the sections. However, when comparing the sections only in terms of the average counts of `AI` triplets, the `walkthrough` section seems to include most, while the prevalence is lower in the sections where the interviewees provide a `basic_job_description` and explain the `dynamic` of their collaboration with AI.

## Similarities to Concept Vectors

To obtain a finer picture of the balance between the user and AI in the collaborative process, we expand the analysis to all triplets extracted from the dataset (limited to the relevant four sections). Additionally, we do not restrict the triplets to the lemma level but obtain a wider span of tokens in the subject and object slots to capture a somewhat larger context and consequently finer semantics (for example, we extract larger phrases like "the AI writes in 500 words", "I double check any facts", "I like to get unique details").

To create reference points to which to compare each of the extracted triplets, we construct two concept vectors: one that would represent user agency (`HUMAN` vector) and one that would represent AI use (`AI` vector). We experiment with two approaches to construct these:

1. concept vectors derived from BERTopic cluster centroids,
2. concept vectors derived from manually labeled triplets from the previous section.

In the second approach, the manually labeled triplets introduced in the previous section serve as the definitive concept anchors. We gather and embed all of their instances using the same sentence encoder as in the previous clustering steps. The manual codebook contains 134 triplets in lemmatized form, of which we find 742 unique forms in the dataset (431 `AI`, 344 `HUMAN`). The reference vectors are then constructed by averaging the embeddings of the triplets belonging to the two labels.

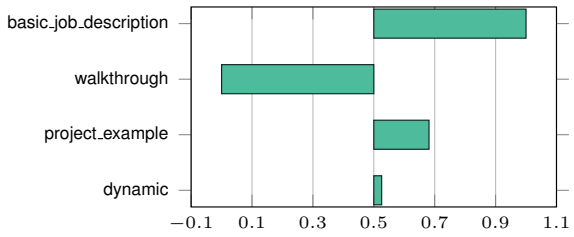


Figure 4: Normalized similarity differences between the 'HUMAN' or 'AI' vector, based on the BERTopic centroid vectors. (The `walkthrough` section leans heavily towards AI use.)

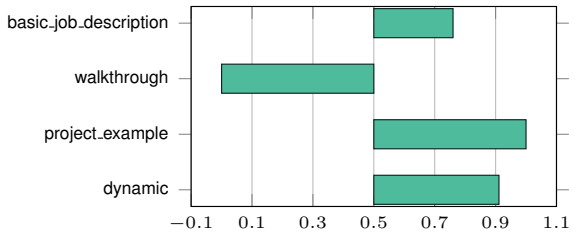


Figure 5: Normalized similarity differences between the 'HUMAN' or 'AI' vector, based on the manually labelled triplets. (The `walkthrough` section leans heavily towards AI use.)

We compute the average similarity of all triplets to the two

concept vectors per section. However, since raw cosine similarities are not directly comparable across the two reference vectors, we derive a normalized difference score.

The **normalized difference** score applies min-max normalization to the raw differences between the HUMAN and AI vectors $\delta(t) = \cos(\mathbf{e}_t, \mathbf{v}_{\text{H}}) - \cos(\mathbf{e}_t, \mathbf{v}_{\text{AI}})$ across the full dataset. This rescales scores to $[0, 1]$, relative to the observed range in the corpus, where 0 indicates maximal proximity to the AI vector and 1 to the HUMAN vector.

$$\text{norm\_diff}(t) = \frac{\cos(\mathbf{e}_t, \mathbf{v}_{\text{H}}) - \cos(\mathbf{e}_t, \mathbf{v}_{\text{AI}}) - \min_{t'} \delta(t')}{\max_{t'} \delta(t') - \min_{t'} \delta(t')} \quad (2)$$

Section-level scores are then obtained by averaging over all triplets within a section. Figures 4 and 5 show the results using the normalized difference score on the whole set of transcripts, and Figure 8 shows the triplet scoring on an interview example. In both BERTopic-based and manual anchor-based cases, clear differences emerge, especially with respect to the `walkthrough` section, which displays the most triplets related to AI use. On the contrary, when looking at the section where users are describing a recent `project example` and the `dynamic` of collaboration, i.e., who is driving the creative process and the decisions therein, we can observe a dominance of triplets related to human agency. This, again, indicates a clear discrepancy between the AI contribution to the creative process and the interviewee's perspective on it.

We also apply our analysis to see whether there are differences between creative types, as we do in the corpus-based and human-judgements scenarios. Figure 6 depicts a heatmap that breaks down the per-section averages depending on the creative job types. From this visualization, it would seem that creative professionals involved in film and video, and those creating web content, use AI the most, while musicians, screenwriters, and photographers seem to use it the least (according to the descriptions in the `walkthrough` section). These results do, in some cases, echo the results from the corpus-based analyses (e.g., highest AI agency for the `walkthrough` section, web-content writers). Nevertheless, we observe that the triplet-level analyses remain incomplete when assessing the exact ratio of AI use vs. human agency. The reasons lie both in the automatic extraction of phrases and their limited context, as well as in limitations inherent to embedding-based approaches, where cosine similarity does not necessarily reflect (only) semantic similarity (see, e.g., Jawahar, Sagot, and Seddah 2019). Most importantly, many of these values are in direct contrast to the judgements provided by human annotators, which we present in the next section.

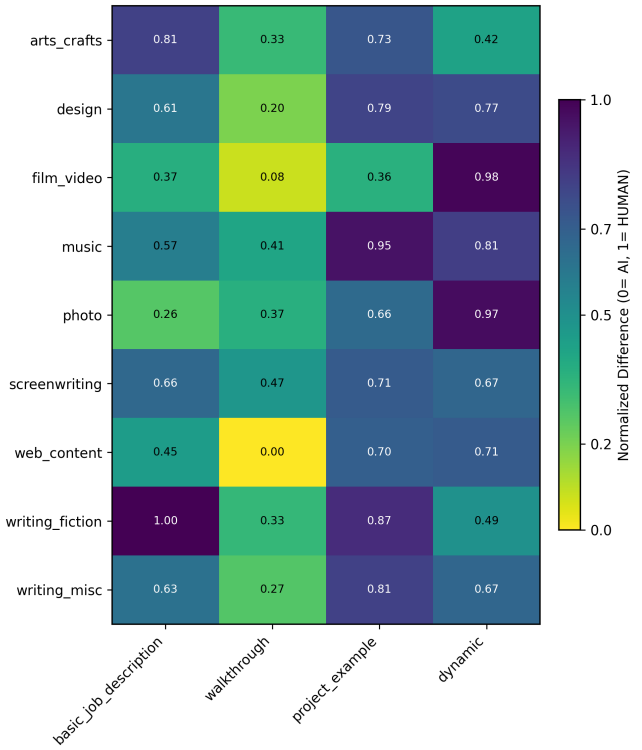


Figure 6: Average section similarities across creative types. (Higher value and darker = HUMAN, lower value and lighter = AI. AI use peaks in the `walkthrough` sections, most visibly among film & video and web content creators.)

## Human Judgements

As each interview describes a unique combination of human skills, expertise, and creativity boosted by various degrees and modes of AI involvement, a fair judgement of AI vs. human input is virtually impossible. We nevertheless devised an experiment where each interview was evaluated by a human creative professional from a similar field. Our group of annotators (N = 9) thus consisted of acclaimed professionals who earn their living through fiction writing, creative writing, music performing and production, video production, graphic and industrial design, photography, and arts and crafts. All but one of them have over 25 years experience in the field. The total set of interviews (N=120) was divided into creative subfields and assigned to annotators, so that no single annotator's workload exceeded 18 interviews.

The annotators were presented only with the first three sections of each interview, labeled Basic Job Description, Use of AI and Recent Project Using AI. The reason for withholding the *Dynamics* section from them was that we wished to obtain independent judgements of AI's contribution without revealing the original authors' self-assessment. The following instructions[8] were provided to them:

[8]The definitions of Ideation and Authorship are based on a rather intuitive, and hence subjective, understanding of the basic features of the creative process, and were formed through informal discussions with creative professionals from different fields. We acknowledge the fact that a different set of instructions might produce different outcomes.

*Below are excerpts from an interview between a human creative professional and an AI agent about the use of AI. Please read it, then assess the contribution of AI to the creative work from two aspects:* **ideation** *and* **authorship**.

***Ideation*** *refers to the creative process and the path from the initial concept or idea to the final execution. This entails generating ideas, thinking and incubation, decision-making and selecting between good and bad options, and includes all the side alleys or subconscious processes running in the background of a creative cycle. In your assessment of AI's contribution, you may consider how you might evaluate the input to ideation if the collaborator were another person.*

***Authorship*** *refers to the creator of an artistic product or project, therefore the person listed as author and the owner of moral copyright*[9]*. Please base your assessment of authorship on the above definition even when assessing products or results not legally protected by copyright. In your assessment of AI's contribution, you may consider how you might attribute authorship if the collaborator were another person.*

The assessments were collected via web forms, where each interview was presented on a separate page with two sliders for ideation/authorship assessment, and a text field for comments.

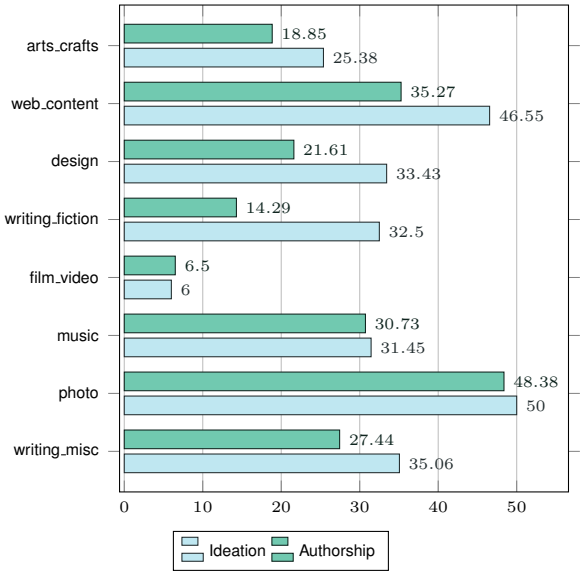


Figure 7: Average human-assessed AI contributions per creative type

According to their comments, most annotators struggled with the task, as assigning a score to someone else's collaborative creation—without actual access to their work or the person—is highly subjective and difficult. Moreover, some transcripts describe more than a single creative project with varying degrees of AI involvement, and the annotators' understanding of the task was individual and domain-specific. For example, the annotator for *design* explained that the AI cannot legally be author, and thus evaluated all AI's authorship shares as 0. The annotator for *music* explained their rationale that the title of a song seemed less part of the song than the actual lyrics, thus the participant who used AI to create song names was considered to be more of an author than someone generating lyrics suggestions.

Figure 7 reveals large differences between creative domains in terms of the human-assessed level of AI-contribution. These may be a reflection of true variation amongst domains due to differences in creative processes or the availability of AI tools, but since each domain was annotated by a single annotator only, we may assume that annotator subjectivity also plays a role. One observation is that ideation scores ($M = 32.44$, $SD = 24.52$) are generally higher than authorship scores ($M = 24.20$, $SD = 25.52$), indicating that AI is perceived as an active collaborator in the conception and idea refinement stages ("I use AI to brainstorm", "I use it to bounce ideas off of", "it helps me out of the creative block"). The lower authorship contribution can also be attributed to the prevailing—and legally canonical—concept of the *human* author; however, the mean value can be misleading: 25% of the transcripts received an AI-authorship score of 40 or more.

## Discussion

Our initial hypothesis was that creative professionals who use AI in their work frequently underestimate AI's contribution to their ideation process, creative decision-making and the final product. Our results weakly support this view, but need to be interpreted carefully.

The limitation of the corpus-based method is its reliance on the manual categorization of only the most frequent triplets, and consequently coverage. Even with a relatively small dataset such as ours, the limited span of the codebook means that, in any single transcript, only a very small portion of the triplets will be found, therefore our scoring of AI vs. human agency is very crude and unable to capture the full span of relevant propositions. Furthermore, the decision to consider `I :: use :: AI` and similar triplets as instances of AI agency is an enforcement of a binary view on the co-creative process, ignoring the nuances which may be present in the context.

The methods based on distributional semantics are more robust in that they expand our search space to a larger set of potentially distinguishing phrases (all triplets containing any of the top 200 verbs), and they tend to reveal the expected trend on the large scale (see Figures 3 and 6). However, we assume that the high proportion of triplets labeled as 'HUMAN' can be attributed to the well-known limitations of measuring semantic relatedness through cosine similarity (see also Jawahar, Sagot, and Seddah 2019): contextual embeddings conflate surface, syntactic and semantic information into a single representation across different layers. Thus, triplets like "I – use – pencil" and "I – use – AI" could be considered similar purely due to surface similarity.

Another important point which needs to be made in relation to all methods and results presented: the human-AI collaborative cases presented in the interviews are varied, complex and unique; and the words and sentences people use to describe their work scenarios, creative processes and attitudes are equally diverse, nuanced and highly individual. While we may successfully identify some patterns in these

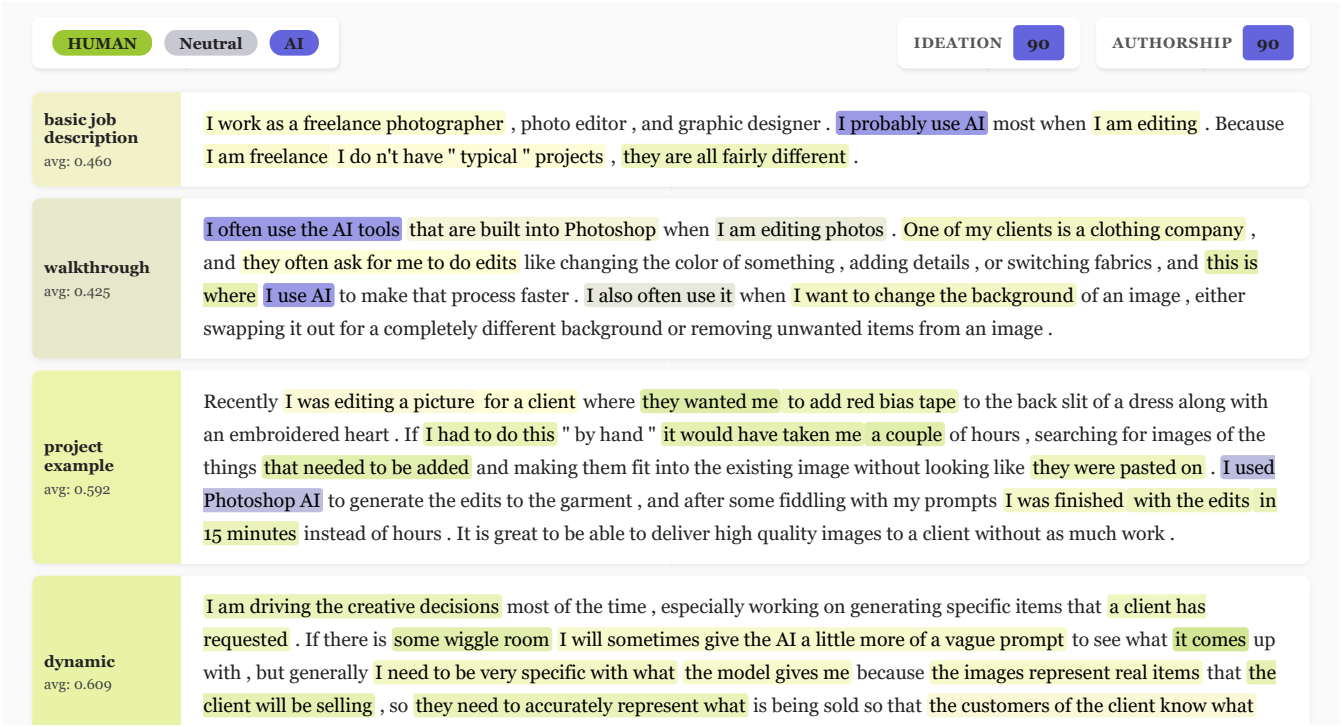


Figure 8: Excerpt from interview transcript. Human scores on the top right. All matching triplets are highlighted by semantic proximity to the concept vectors in terms of the normalized difference score. Green indicates closer proximity to the `HUMAN` vector, and blue closer proximity to the `AI` vector. The example shows the large discrepancy between the human scores for AI share in Ideation and Authorship and the interviewee's description of the collaboration in the fourth section (`dynamic`).

accounts, we cannot claim to be modeling the full complexity of the human-AI creative expression.

Still, at least for the use-cases which involve large language models, the idea that humans overrely on the LLM's overconfident responses while maintaining the illusion of their own control has been addressed - and proven - beyond the scope of computational creativity (Rathi, Jurafsky, and Zhou 2025), and the so-called *fictional human oversight* has become a known issue in debates on AI safety (Kasneci and Kasneci 2026). We argue that in the era of LLMs and AI-powered generative platforms the traditional models of human-computer co-creativity need to be revised to accommodate - and monitor - the challenges arising from the "unbearable lightness of AI-ing", and that raising the awareness about these challenges amongs creative professionals could help relieve the pressure that many of them feel in the race between humans and AI.

## Conclusions

We present an experiment to identify potential egocentric bias in human-AI collaborative scenarios, i.e., the tendency to claim full control over the creative process, and to overestimate one's own contributions over those of the AI. We approach the problem with a set of linguistic methods, and perform human assessment for comparison.

The first method is traditional corpus-based analysis supported by manual annotation of a sample of prominent triplets, which might indicate human vs. AI agency. While the resulting AI-agency scores successfully show the cumulative shifting perspectives of the interviewees through the interview sections, it cannot be considered accurate enough due to low coverage. The set of methods based on distributional semantics seem more promising, although results also indicate that cosine similarity may blur true semantic proximity with surface-level phenomena. Finally, human assessments provide an insight into how other creative professionals perceive AI's contribution to the ideation stage and the authorship of the final product.

Our results tentatively support the initial hypothesis, and at the same time reveal the complexity of the human-AI collaboration dynamics. While the methods presented could certainly be improved in the future, we believe that the main contribution of this paper lies in the topic itself, and the associated ethical, legal, psychological and philosophical dimensions of authorship attribution in the era of GenAI. A discussion of these dimensions lies beyond the scope of our work (or our expertise), but has been and will remain a recurring issue in the creative community. We finally emphasize that GenAI was not used to produce any part of this paper, nor was it used in the analysis of the data or interpretation of the findings. It was however used in the research of related work, and as a coding aid. Are we still driving?

## Author Contributions

Author 1 was responsible for planning the study, designing the human judgements experiment and synthesizing results, Author 2 was responsible for dataset preparation and the experimental part using clustering and semantic distance, Author 3 was in charge of the corpus-based analysis and Author 4 for the recruitment of creative professionals and visualizations of results. All authors contributed to the writing of this manuscript.

## Acknowledgments

This work was supported by the projects "Large Language Models for Digital Humanities" (Grant GC-0002), "Slovene Language - Basic, Contrastive, and Applied Studies" (P6-0215), the "Jožef Stefan" Infrastructure Programme (Grant I0-0005) and the Slovenian Artificial Intelligence Factory project (SLAIF).

The authors sincerely thank the creative professionals who participated as annotators in this study.

## References

[Abbott and Rothman 2022] Abbott, R., and Rothman, E. 2022. Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence. doi:10.2139/ssrn.4185327.

[Boden 1992] Boden, M. 1992. *The Creative Mind*. London: Abacus.

[Boo, Kim, and Suh 2025] Boo, C.; Kim, Y.; and Suh, A. 2025. A Collaborative Creative Process in the Age of AI: A Comparative Analysis of Machine and Human Creativity. In *Proceedings of the 58th Hawaii International Conference on System Sciences*. doi:10.24251/HICSS.2025.025.

[Bown et al. 2020] Bown, O.; Grace, K.; Bray, L.; and Ventura, D. 2020. A Speculative Exploration of the Role of Dialogue in Human-Computer Co-creation. In *Proceedings of the International Conference on Computational Creativity 2020*.

[Breidenthal et al. 2020] Breidenthal, A. P.; Liu, D.; Bai, Y.; and Mao, Y. 2020. The dark side of creativity: Coworker envy and ostracism as a response to employee creativity. *Organizational Behavior and Human Decision Processes* 161:242–254. doi:https://doi.org/10.1016/j.obhdp.2020.08.001.

[Condorelli and Berti 2025] Condorelli, F., and Berti, F. 2025. Creativity and awareness in co-creation of art using artificial intelligence-based systems in heritage education. *Heritage* 8(5). doi:10.3390/heritage8050157.

[Dunning 1993] Dunning, T. E. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1):61–74.

[Ginsburg and Budiardjo 2018] Ginsburg, J. C., and Budiardjo, L. A. 2018. Authors and machines. *Berkeley Technology Law Journal* 34:343. `https://api.semanticscholar.org/CorpusID:69352843`.

[Grootendorst 2022] Grootendorst, M. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

[Gur and Sackeim 1979] Gur, R. C., and Sackeim, H. A. 1979. Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology* 37(2):147–169. doi:10.1037/0022-3514.37.2.147.

[Handa et al. 2025] Handa, K.; Stern, M.; Huang, S.; Hong, J.; Durmus, E.; McCain, M.; Yun, G.; Alt, A.; Millar, T.; Tamkin, A.; Leibrock, J.; Ritchie, S.; and Ganguli, D. 2025. Introducing Anthropic Interviewer: What 1,250 professionals told us about working with AI. `https://anthropic.com/research/anthropic-interviewer`. `https://anthropic.com/research/anthropic-interviewer`.

[Hardie 2014] Hardie, A. 2014. Log ratio–an informal introduction. *ESRC Centre for Corpus Approaches to Social Science (CASS)* 1–2.

[Jawahar, Sagot, and Seddah 2019] Jawahar, G.; Sagot, B.; and Seddah, D. 2019. What does BERT learn about the structure of language? In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1356.

[Kantosalo and Toivonen 2016] 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.

[Karimi et al. 2018] Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating Creativity in Computational Co-Creative Systems. doi:10.48550/arXiv.1807.09886.

[Kasneci and Kasneci 2026] Kasneci, G., and Kasneci, E. 2026. The safety failures we are not instrumenting: A perspective on hidden safety-critical challenges in modern AI systems. *AI and Ethics* 6(3):295. doi:10.1007/s43681-026-01132-0.

[Kilgarriff and Tugwell 2001] Kilgarriff, A., and Tugwell, D. 2001. Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of the ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, 32–38.

[Kilgarriff 2001] Kilgarriff, A. 2001. Comparing corpora. *International journal of corpus linguistics* 6(1):97–133.

[Liu and Huang 2025] Liu, W., and Huang, W. 2025. Authorship in Human-AI collaborative creation: A creative control theory perspective. *Computer Law & Security Review* 57:106139. doi:10.1016/j.clsr.2025.106139.

[Mele 2001a] Mele, A. R. 2001a. *Autonomous Agents: From Self-control to Autonomy*. Oxford University Press.

[Mele 2001b] Mele, A. R. 2001b. *Self-Deception Unmasked*. Princeton University Press.

[Nikrang and Kiesenhofer 2025] Nikrang, A., and Kiesenhofer, S. 2025. Human–AI co-creation in contemporary composition: Interaction and artistic strategies with Ricercar. In *Proceedings of the Conference on Animation and Interactive Art*, Expanded '25, 65–73. New

York, NY, USA: Association for Computing Machinery. doi:10.1145/3749893.3749961.

[Pruschak and Hopp 2022] Pruschak, G., and Hopp, C. 2022. And the credit goes to ... - ghost and honorary authorship among social scientists. *PLOS ONE* 17(5):1–22. doi:10.1371/journal.pone.0267312.

[Rathi, Jurafsky, and Zhou 2025] Rathi, N.; Jurafsky, D.; and Zhou, K. 2025. Humans overrely on overconfident language models, across languages. In *Second Conference on Language Modeling*. https://openreview.net/forum?id=QsQatTzATT.

[Ross and Sicoly 1982] Ross, M., and Sicoly, F. 1982. *Egocentric biases in availability and attribution*. Cambridge University Press. 179–189.

[Rossiter 1993] Rossiter, M. W. 1993. The Matthew Matilda effect in science. *Social Studies of Science* 23(2):325–341. http://www.jstor.org/stable/285482.