International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli

Simone Grassini & Mika Koivisto

To cite this article: Simone Grassini & Mika Koivisto (02 May 2024): Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2024.2345430

To link to this article: https://doi.org/10.1080/10447318.2024.2345430

Published online: 02 May 2024.

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2024.2345430

Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli

Simone Grassini[a,b] and Mika Koivisto[c]

[a]Department of Psychosocial Science, University of Bergen, Bergen, Norway; [b]Cognitive and Behavioural Neuroscience Laboratory, University of Stavanger, Stavanger, Norway; [c]Department of Psychology, University of Turku, Turku, Finland

ABSTRACT

This study investigates the creative capacities of artificial intelligence (AI), exemplified by ChatGPT-4, in comparison with human creativity, utilizing the Figural Interpretation Quest (FIQ) as the evaluative tool. The primary objective is to assess whether AI can surpass human performance in creative tasks, particularly in terms of flexibility and subjectively perceived creativity. Participants, including both ChatGPT-4 and humans, were asked to provide creative interpretations for a set of ambiguous, abstract figures. The study measured the flexibility of these interpretations and gathered subjective ratings of creativity from independent evaluators. Results indicated that while AI, on average, demonstrated higher flexibility in generating creative interpretations, human participants excelled in terms of subjectively perceived creativity. Furthermore, the most creative human responses were higher than those of AI in both flexibility and subjective creativity. These findings suggest that in these types of complex multimodal interpretation tasks, AI shows excellent potential and the ability to generate semantically diverse ideas, yet human subjectively perceived creativity remains for now unmatched. The study highlights the limitations of AI in replicating the complexity of human creativity and points to possible theoretical interpretations of the reported findings.

KEYWORDS

AI; artificial intelligence; creativity; divergent thinking

1. Introduction

The advent of effective and user-oriented generative artificial intelligence applications, like ChatGPT by OpenAI, Llama from Meta, and Google Bard, has catalyzed an extensive and multifaceted dialogue about their societal impact (Fui-Hoon Nah et al., 2023) and has prompted a reevaluation of the concept of creativity within human and AI domains (Millet et al., 2023). Recent developments of AI systems have prompted discourse regarding the potential repercussions employment sectors (Kim, 2023). As AI demonstrates to gain proficiency for tasks traditionally reserved for humans, the displacement of human labor and the subsequent implications for the workforce of tomorrow have also emerged (Anantrasirichai & Bull, 2021).

In the context of education, the proliferation of AI tools raises pedagogical questions on how the use of these tools may be used by both teachers and students (Grassini, 2023) and whether these tools should be banned from academia (Yu, 2023). Furthermore, it has been discussed that the use of AI tools may potentially cause a decline in critical thinking skills among learners (Spector & Ma, 2019). Additionally, the generation of content by AI has expanded the scope of the debate to include legal and ethical dimensions (Hacker, 2021; Prem, 2023). As AI becomes more sophisticated at content creation, the developers of such generative AI systems have come under legal scrutiny, facing copyright infringement claims from original content creators (Samuelson, 2023).

The emergence of machines capable of producing creative and informative outputs, such as text and visual art, provokes a rethinking of the dynamics between human and machine interaction, particularly with regards to the future landscape of creative expression (Mazzone & Elgammal, 2019). While AI technologies have succeeded in automating segments of the creative process, critiques have emerged suggesting that such automation often yields products lacking in genuine originality. This perceived shortcoming is often linked to the dependency of AI on pre-existing datasets, which may inherently limit its potential for true novelty and unique creation (Lee, 2011).

Nevertheless, recent studies have highlighted the proficiency of AI in completing tasks that are traditionally used to measure human creativity, suggesting a closer parity between human and machine creativity than previously acknowledged (Guzik et al., 2023; Haase & Hanel, 2023; Koivisto & Grassini, 2023). Recent studies have found that AI artistic production is found undistinguishable from human creative endeavor (Hitsuwari et al., 2023). Additionally, research indicates that AI can engage in genuine artistic endeavors that can be perceived as being beyond mere replication or reuse of the training data (Arriagada, 2020; Carnovalini & Rodà, 2020). This suggests a potential shift in the understanding of creativity as a characteristic that might not be

CONTACT Simone Grassini simone.grassini@uib.no Department of Psychosocial Science, University of Bergen, Christies gate 12, 5015 Bergen, Norway

exclusively human and raises questions about the future of innovation in an increasingly AI-integrated world.

These debates regard existential questions about what it means to be humans and the nature of skills as creativity that have historically always been associated with human being (Millet et al., 2023). As AI technology advances, it challenges entrenched perceptions of human identity and prompts a reassessment of the attributes that distinguish our species. The concept of creativity, traditionally ascribed solely to the conscious endeavor of humans (Arriagada, 2020; Chamberlain et al., 2018), is now being critically analyzed in the context of AI's ability to generate what is perceived as original content.

Artificial intelligence has unveiled vast and expanding horizons in domains that demand analytical reasoning and inventive decision-making capabilities. An example of this is the evolution of advanced chess engines (Duca Iliescu, 2020). These AI systems have not only challenged but also triumphed over chess experts, redefining the game's competitive landscape (Wilkenfeld, 2019). AI systems have become extremely good at beating humans in other games that require analytical reasoning such as GO (Binder, 2020) and other complex board games (Purves, 2019).

Moreover, AI's venture into the realm of artistic creation has been marked by notable success. AI-powered tools have demonstrated their ability to craft art pieces of excellent aesthetic quality that cannot be distinguished from human creations (Oksanen et al., 2023). This breakthrough in generating artwork through AI disrupts traditional notions of artistry and that AI may begin to carve out a new niche in the creative industries, one where machine intelligence plays a significant role in the conception and realization of art.

It is difficult to provide a general definition for creativity as a psychological construct (Mehta & Dahl, 2018). The concept of creativity is often defined as the capacity to generate ideas that are both novel and practical, a definition first articulated by Guilford in 1967 (Guilford, 1967). This definition enables the assessment of AI-generated ideas using the same standards as those for human ideas. Our study examines the creative outputs of both AI and humans, drawing on Guilford's distinction between convergent and divergent thinking. Convergent thinking entails finding the best or correct solution to a problem, while divergent thinking involves producing a variety of ideas or solutions. Divergent thinking is more closely linked to creativity, representing the ability to imagine multiple potential answers. It includes aspects like fluency, flexibility, originality, and elaboration (Silvia et al., 2008). However, the notion that divergent thinking is synonymous with creativity has been debated (Silvia et al., 2008). Despite this, divergent thinking remains a key focus in psychological research on creativity, with established assessment tasks (Silvia et al., 2008).

Flexibility is often referred to as a component of creativity, serving as a dynamic mechanism that supports creative processes (Cho et al., 2010). Flexibility is often defined as the ability to produce a diverse array of ideas or solutions to a problem (Clément, 2022; Georgsdottir & Getz, 2004). In the domain of psychological research, particularly in the study of creative thinking, the significance of flexibility is paramount. This mental faculty, as emphasized in several scientific works (see e.g. Chi, 1997; Jaušovec, 1991, 1994; Runco & Okuda, 1991), is central to fostering creativity. Cognitive flexibility is characterized by the capacity to deviate from established cognitive pathways, surmount the barriers of functional fixedness, and consequently, forge innovative associations among various concepts (Guilford, 1967).

Creative process theories often adopt a dual-process view, with Guilford's (Guilford, 1967) model positing that creativity involves a combination of spontaneous (divergent) and controlled (convergent) thinking. Spontaneous thinking drives the originality of ideas, while controlled thinking evaluates their applicability. The associative theory of creativity (Kenett & Faust, 2019; Mednick, 1962) suggests that creativity stems from linking weakly related concepts to create novel ideas. This theory is supported by computational methods (Kenett & Faust, 2019) and brain imaging studies (Ovando-Tellez et al., 2022), indicating that creative individuals possess more connected and adaptable semantic networks.

The controlled-attention theory, highlighted by Beaty and colleagues (Beaty et al., 2014), underscores the importance of executive functions in generating creative ideas. This theory can be merged into a hybrid view that combines bottom–up, associative processes and top–down executive control during concept retrieval from semantic memory. For example, top–down processes help in generating and maintaining retrieval cues, suppressing obvious associations, and shifting focus (Beaty et al., 2014).

1.1. The present study

Divergent thinking is typically measured through tests that require open-ended responses, such as the AUT, as defined in Guilford (1967), and such task was used already to compare human and AI creativity (Koivisto & Grassini, 2023). In the present study, we used the newly developed Figural Interpretation Quest (FIQ) to evaluate divergent thinking (Erwin et al., 2022). The FIQ is a perceptually based multimodal task that involves not only text processing, but also visual inputs. The task challenges the participants to provide creative interpretations for ambiguous, abstract figures. The FIQ and its relationship with other measures of creativity have been examined in the literature (Erwin et al., 2022; Koutstaal et al., 2022). This task assesses creativity through divergent thinking, specifically evaluating the originality, fluency, and variety of ideas generated in response to ambiguous stimuli. This measure complements traditional text-based divergent thinking tasks like the Alternative Uses Task (AUT) and has shown significant predictive validity for originality in domain-specific idea generation, such as in Design Product Ideation tasks (Erwin et al., 2022). In evaluating the FIQ's validity and relationship with other creativity measures, studies have found that while FIQ is related to other divergent thinking tasks, it does not measure identical constructs. These findings emphasize the importance of incorporating both conceptual and perceptual divergent thinking measures in assessing creativity (Erwin et al., 2022). Furthermore, compared to other text-based tasks, performing the FIQ may require some level of mental imagery (Finke, 1996). Historically, various authors (see e.g. González et al., 1997) have proposed a link between mental

imagery and creativity. Mental imagery is a common representation technique in daily life (Antonietti & Colombo, 1997) significantly aiding in thinking. These images serve two primary roles: as cognitive simulation tools and as symbolization mediums (Kosslyn, 1980). In their role as simulators, mental images allow individuals to mentally rehearse actual operations and physical changes, maintaining an analogic relationship with the external world. This internal representation is invaluable for visualizing outcomes that might not be immediately apparent in verbal or abstract forms. As symbols, they represent objects or events through conventional signs, reducing memory load and facilitating smoother, faster manipulation of elements (Antonietti, 1991; Antonietti & Colombo, 2011; Kosslyn, 1980).

Regarding creativity, mental images enhance the process of identifying similarities, differences, and connections between diverse elements—key components of creative thinking (Antonietti et al., 2020). Mental images may also allow for the reorganization of one's perception of a situation, leading to more productive considerations. Importantly, they provide a visual form for abstract concepts, allowing for the simultaneous representation of various elements and the discernment of their interrelations. Thus, mental images are particularly effective in the creative process, owing to their flexibility, ease of transformation, and ability to amalgamate multiple elements into novel concepts (Antonietti, 1991).

Several scores for flexibility using automatic computational (text-mining-based methods) approaches have been recently proposed (Grajzel et al., 2023). In our study, we employed the SemDis platform proposed by Beaty and Johnson (2021). This method was used to obtain an objective score to assess the semantic distance between two interpretations of the same figure of the FIQ. The larger the distance between two interpretations, the more flexibility is shown. For example, interpretations of a FIQ figure belonging to the same category typically get a lower semantic distance score (e.g. the mean score computed with SemDis across five semantic spaces is 0.46205 for "hair-wig") than responses belonging to different semantic categories (0.9583 for "hair-boomerang"), indicating more flexibility in interpreting the stimulus from different perspectives. The FIQ was developed recently, and therefore, ChatGPT-4 should not have had access to optimal responses to the task in its training data (ChatGPT-4 release 27.09.2023 was used during data collection in October 2023, when the tests were made, ChatGPT-4 training data was allegedly limited up to September 2021 according with ChatGPT-4 itself). In addition, eleven blinded human evaluators judged the responses, offering a subjective perspective on creativity, recognizing that semantic distance alone might not fully capture creativity, but only flexibility in the context of the proposed task.

Recent research that has compared AI and Human creative outputs (Guzik et al., 2023; Haase & Hanel, 2023; Koivisto & Grassini, 2023), has found that modern chatbots as ChatGPT have reached human-level creativity level, even surpassing the average human responders in psychological tasks commonly used to assess human creativity as the AUT (Haase & Hanel, 2023; Koivisto & Grassini, 2023), or the Torrance test (Guzik et al., 2023). However, all these studies have shown that the best performers among humans are still able to outperform even the most modern of the AI systems. One limitation of these studies is the use of relatively old tests for assessing creativity of which a significant quantity of materials and examples could be found online. Such materials may have been used as training data for the chatbot that would then repropose when prompted with the task. In this case, the creative results of the chatbots would not reflect true creative behaviors but a memory output from the stored training data. Another limitation of the previous studies is that all used text-based prompt and responses, without involving interpretative or multimodal effort; however, human creativity is often displayed in multimodal contexts (Finke, 1996; Riquelme, 2002).

It is essential to highlight the distinct challenges and requirements of verbal versus visual creative tasks in AI. The verbal AUT, for instance, assesses creativity by asking participants to think of as many uses as possible for a common object, thereby measuring divergent thinking through linguistic expression. On the other hand, tasks like the FIQ explore the ability to manipulate and transform visual patterns and shapes, an ability that may be related to visualization and mental imagery. Visual pattern recognition and interpretation have traditionally been challenging areas in machine learning primarily due to the complexity and variability of visual data compared to textual information (LeCun et al., 2015). The distinct nature of these tasks underscores the importance of understanding the varying capabilities of AI in different creative domains. As AI continues to evolve, exploring and enhancing its proficiency in multimodal creative tasks, including those requiring visual imagination and interpretation, becomes increasingly relevant.

Assuming the results from previous studies on verbal divergent thinking generalize to divergent thinking tasks requiring interpretation of ambiguous visual stimuli (Guzik et al., 2023; Haase & Hanel, 2023; Koivisto & Grassini, 2023), we had preregistered the following hypotheses:

H1: On average, AI performs better than humans in the FIQ task.

H2: The best humans' creative responses are better than any AI output.

Exploratory analyses (H3): There is a strong positive correlation between flexibility as measured with semantic distance and creativity measured with humans' subjective ratings.

2. Method

The study hypotheses and data analysis process were pre-registered using Open Science Framework (https://osf.io/xv3n2/?view_only=1d9a96705c3c4b79aa9207ba52295ee9). Deviations from the pre-registration are explicitly mentioned in the article text.

All methods were carried out in accordance with the relevant guidelines and regulations. According to national and local regulations of the country/institution, further approval of the experimental protocol was not necessary for this study.

2.1. Participants

Native English speakers' participants were enlisted through Prolific's online platform (www.prolific.co) for a 13-minute study, receiving £2 as compensation. The study was accomplished using Psytoolkit (Stoet, 2010, 2016). Out of 310 individuals who accessed the study link, complete data were acquired from 279 who completed it fully. Among these, 256 (comprising 108 females, 145 males, 2 identifying as other, and 1 opting not to disclose their gender) successfully passed attention checks involving simple visual tasks, and their data were utilized in the study. The participants had an average age of 30.4 years, spanning from 19 to 40 years, with 44 being students. Employment status was: 142 were full-time, 37 part-time, 30 unemployed, and 42 in other conditions like homemaking or retirement. They reported no significant medical issues and resided in the UK (166), USA (79), Canada (9), or Ireland (2). The human data were gathered after participants' informed consent and the procedure followed the principles of the Declaration of Helsinki. The online platforms used in the study met the standards of the European and UK data protection law (the General Data Protection Regulation, GDPR; https://prolific.notion.site/Privacy-and-Legal-at-Prolific-395a0b3414cd4d84a2557566256e3d58; https://www.psytoolkit.org/faq.html#ethics). The study was anonymous, and no sensitive or direct personal data were collected from the participants.

The AI chatbot ChatGPT-4 was tested between 19.10.2023 and 23.10.2023. The chatbot was tested 247 times using four figures of the FIQ as prompts in different sessions. The number of sessions reflected the number of valid participants in the study with humans, after participants with one or more empty responses and participants with at least rarely long response (more than three words) were removed (respectively 1.18% and 0.98% of all participants in average across the four FIQ tasks). The chatbot was asked to produce responses of 1-, 2-, or 3-words length without considering common English stop words, to somehow imitate the characteristic responses given by human participants. This distribution of response lengths was established to reflect the distribution of response length from the human data. As ChatGPT-4 allegedly does not have memory between the various chat sessions, the chat was restarted for every session. The prompt with the instruction and the upload of the four images in the ChatGPT-4 chat were then iterated for each session.

The participant type (human vs. AI) was the independent variable, while the measured dependent variables included the flexibility of divergent thinking (assessed via semantic distance) and subjectively perceived creativity (assessed by human raters).

2.2. Stimuli and procedure

Four different ambiguous figures from the FIQ stimulus set (Erwin et al., 2022) were used in the task presented to both humans and ChatGPT-4. The figures were always the same and were presented in the same order to both human participants and ChatGPT sessions. For examples of the used figures, see Figure 1. Please note that among the examples reported in Figure 1, only the first figure on the left was used in our study, while the other two figures represent examples of FIQ items previously openly published in papers, see Erwin et al. (2022). The other FIQ figures that we used in the present study cannot be openly shared in a publication as for agreement with the FIQ test developers (Erwin et al., 2022; Koutstaal et al., 2003), as making them publicly available may invalidate the use of the task in future studies[1]. The figures included in the FIQ (Erwin et al., 2022) were a selection of those used in an earlier study (Koutstaal et al., 2003).

The human participants were asked to give two interpretations to the ambiguous figures, with a focus on providing two interpretations that were as semantically different as possible. An example figure and interpretation were also presented to the participants, alongside a short description of what semantic distance means. They were asked to "Come up with two interpretations for what the image could represent. Think creatively! Try to invite two as different interpretations as you can." They were then shown four ambiguous images, one at the time, and they were to write their interpretations into two textboxes displayed in the screen below the image. Human participants were given 30 seconds to enter the two interpretations for each of the figures. The figures always appeared in the same order for all participants.

The new feature of image processing introduced for plus users in October 2023 was used to provide ChatGPT-4 with the figures of the FIQ. ChatGPT-4 was given the same images and instructions as human participants. Differently than for human participants, ChatGPT-4 was not given the sample image, as we assumed that the chatbot could operationalize what semantic distance is, as it is a common concept often mentioned in the scientific literature, and therefore, the chatbot was assumed to contain such

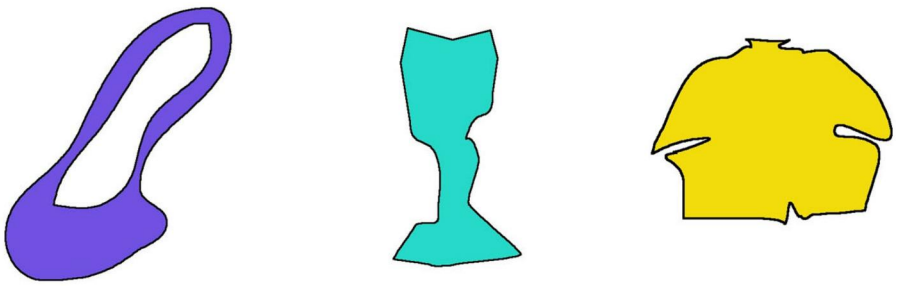

Figure 1. Three sample items from the Figural Interpretation Quest (Erwin et al., 2022). Reproduced with permission. Please note that, among the sample items here presented, only the first figure on the left has been used in the present study.

information in its training data. The AI was prompted to generate creative interpretations using the same task description as used for human participant, however with an established (variable) word limits for a number times in different chat sessions (either 1, 2, or 3 words). This word limit was given to obtain the same word-length distribution that preliminary analysis of human data had shown.

Human participant data were collected as part of another study, and such data collection was previously pre-registered at OSF.io (https://osf.io/fy3mn/?view_only=f1cf960d0170433dba9d31df68a6eaf7). The AI data were collected in October 2023 using ChatGPT-4 with plus subscription.

2.3. Scoring

To assess these variables, two main indices were used:

Flexibility of Divergent Thinking: Measured using the semantic distance between the two interpretations for each image (Grajzel et al., 2023), calculated via the SemDis platform (semdis.wlu.psu.edu). In the SemDis semantic distance analysis, the "multiplicative" compositional model was utilized to handle AUT responses consisting of several words. These responses underwent preprocessing with the "remove filler and clean" options which eliminates common English "stop words" such as "the," "an," "a," "to," and punctuation, to avoid affecting the computation of semantic distance. For each pair of two interpretations of a figure, the mean semantic distance was computed across five semantic spaces (for a detailed description, see Beaty & Johnson, 2021).

Subjectively Perceived Creativity: The raters were given the task to evaluate every proposed interpretation of the images according to the perceived level of creativity using a scale with three alternatives. 0: Either a basic/non-creative interpretation or an interpretation seemingly unrelated to the Figure 1: A creative take on the Figure 2: An exceptionally creative interpretation, notable for its originality. A 0 to 2 scoring scale was previously used to evaluate the interpretation of the FIQ (Erwin et al., 2022). The raters were asked to approach each interpretation with an open mind, recognizing the several different ways the figure might be perceived by different participants. The interpretations were presented in four separate sheets, one for each figure, and the order of the interpretations within each sheet was randomized separately for each rater. The raters were psychology students from the University of Bergen (Norway) or the University of Turku (Finland), and they were blind to the origin of the texts they evaluated. They conducted this scoring as part of their research training or in exchange for university credits. We initially asked three people to perform the task (pre-registered scoring procedure); however, the data from these three raters showed a weak interrater reliability (ICC), of 0.4, 0.3, 0.3, 0.15, for the four FIQ figures respectively. Therefore, we recruited an additional 8 raters, for a total of 11 raters to perform the rating of the interpretations in the attempt to make the overall evaluations more reliable. The ICC for each of the FIQ figures considering the 11 raters were: ICC1 = 0.702, 95% CI [0.662 < ICC < 0.74], ICC2 = 0.415, 95% CI [0.335 < ICC < 0.489], ICC3 = 0.354, 95% CI [0.266 < ICC < 0.435], ICC4 = 0.75, 95% CI [0.716 < ICC < 0.781]. These ICC showed an improved interrater reliability, especially for the rating of the interpretations of the FIQ 1 and 4.

2.4. Statistical analyses

Linear mixed-effect analyses were performed with lme4 package (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) in R (R Core Team, 2018) on the single trials (each single interpretation in the analysis of creativity; each single semantic distance score in the analysis of flexibility). In the first set of models, the only fixed effect was Group (GTP-4 vs. human) and random intercept for participants (and session for AI) served as the random effect. In the next set of analyses performed on creativity and flexibility, Group and Figure, and their interaction were the fixed effects. In these analyses the R's ANOVA function was applied on the models to obtain Type III analysis of variance results (Satterthwaite's method). The contrasts comparing the groups separately in response to each figure were performed with the package emmeans v.1.8.2. (https://CRAN.R-project.org/package=emmeans).

3. Results

In the preliminary analyses, the length of the responses (i.e. the number of non-stops words in each interpretation) from

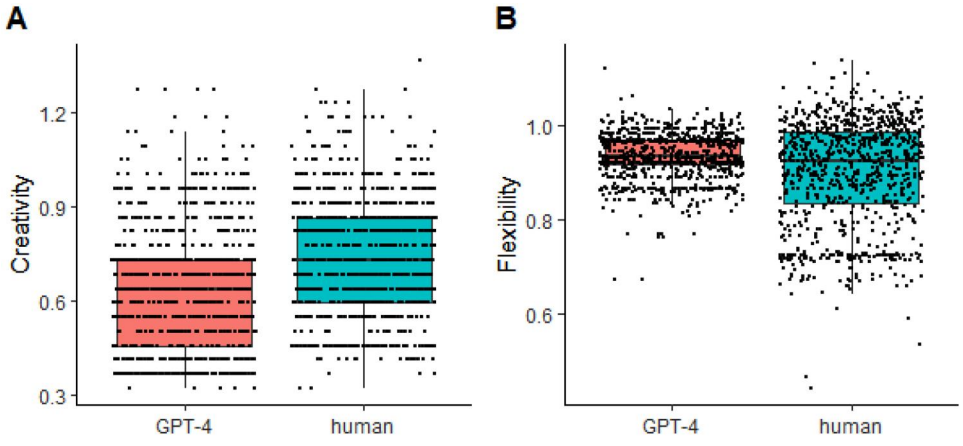

Figure 2. GPT-4's and humans' (A) creativity scores and (B) flexibility scores across all trials.

humans were analyzed. This was done to be able to replicate, via prompt, the same distribution of response length by the chatbot (see section 2.3). We purposely wanted to limit the response length from the chatbot: as the machine being generally faster in writing than a human, would have surpassed humans in terms of response length, possibly affecting the perceived level of creativity of the responses. The preliminary analyses revealed that 1.18% of human responses were of 0 words (empty), 76.32% were of 1 word, 18.43% were of 2 words, 3.09% were of 3 words, and 0.98% were of more than 3 words. Therefore, we only considered humans that did not have empty responses (since we judged humans that did not delivered the two responses as not concentrated enough or inattentive to the task), and very rare long responses (responses longer than three words).

The rest of the analyses were performed with all valid participant data. Please note that human participants dropped from 247 to 246 after calculating the flexibility rating using the semdis, as it was not possible by the algorithm to calculate the score for one of the participants. Data from the remaining 246 participants were analyzed according to the pre-registered data analysis protocol.

Some examples of response pairs that were given for the FIQ figures are presented below. Within squared parentheses is reported the flexibility score based on sematic distance as calculated using the SemDis method. The responses were given in response to the FIQ figure shown on the left in Figure 1. The examples are taken among the highest and the lowest scored pairs. Examples from human participants: low score, "necklace – bracelet" [0.3727] and "rope-scarf" [0.7515]; high score, "screaming hooded figure – lake with island" [1.05766]. Examples from ChatGPT: low score, "bean – ladle" [0.76815]; high score "music note – water droplet" [1.05304].

3.1. Descriptive statistics and correlations

Table 1 shows the descriptive statistics for GPT-4 and humans, averaged across all the four FIQ objects. The correlation between of creativity (human-made subjective ratings) and flexibility (semantic distance between responses) was weak in the human group (r = .175, 95% CI [.051, .294], p = .013), and no correlation was detected between GPT-4's creativity and flexibility, r = .071, 95% CI [−.054, .194], p = .498). Three humans (1.2%) scored in total creativity score (i.e. mean of all interpretations to the 4 FIQs) better than the best GPT-4 session total score, and 13 humans (5.3%) scored better than the best GPT-4 session in total flexibility. In total, humans produced 794 unique responses to the task, while ChatGPT-4 produced only 385 responses, showing an overall smaller range of responses for AI compared to humans.

Overall performance

The overall differences between GPT-4 and humans, across the four figures, were analyzed with linear mixed-effect models with Group (GPT-4, human) as the fixed effect. Overall, humans' creativity scores (Figure 2(A)) were higher than those of GPT-4, B = 0.107, SE = 0.008, 95% CI [0.090, 0.124], t(1946) = 12.64, p < .001. However, flexibility (i.e. semantic distance between the responses) (Figure 2(B)) was higher for GPT-4 than humans, B = −0.036, SE = 0.004, 95% CI [-0.044, −0.027], t(484.04) = −8.493, p < .001.

Creativity

Next, we studied in more detail the creativity of the responses of GPT-4 and humans to each figure, with Group (GPT-4, human) and Figure (the 4 FIQs) and their interaction as fixed effects (Figure 3). The main effects for Group, F(1, 489.79) = 222.23, p < 0.001, η_p^2 = .31, and Figure, F(3, 1459.26) = 198.386 < 0.001, η_p^2 = .29, were statistically significant. In addition, the Group x Figure interaction was significant, F(3, 1459.26) = 77.86, p < 0.001, η_p^2 = .14. The contrasts showed that humans' responses were more creative than those of GPT-4 to FIQ1, B = −0.086, SE = 0.014, 95% CI [-0.114, −0.058], t(1939) = −6.095, p < .001, to FIQ2, B = −0.106, SE = 0.014, 95% CI [−0.134, −0.078], t(1939) = −7.434, p <.001, and to FIQ4, B = −0.272, SE = 0.014, 95% CI [−0.300, −0.244], t(1939) = −19.094, p < .001. The responses of GPT-4 to FIQ4 were more creative than those of humans, B = 0.031, SE = 0.014, 95% CI [0.003, 0.059], t(1939) = 2.177, p = .0296.

Flexibility

GPT-4 showed more flexibility than humans, F(1, 483.96) = 69.33, p < .001, η_p^2 = .13. Unlike creativity, flexibility on average was higher for GPT-4 than humans (Figure 4). The main effect of Figure, F(3, 1447.45) = 23.91, p < .001, η_p^2 = .05, and the Group × Figure interaction, F(3, 1447.45) = 37.35, p < .001, ηp2 = .07, were statistically significant, suggesting that the superiority of GPT-4 did not generalize to all figures. Contrasts showed that GTP-4's scores were larger than humans' in response to FIQ1, B = 0.084, SE = 0.007, 95% CI 0.070, 0.098], t(1882) = 11.42, p < .001, and to

Table 1. Descriptive statistics for creativity (human-made subjective ratings) and flexibility (semantic distance between the responses), averaged across all responses to the four figures.

				95% confidence interval				
	Group	*N*	Mean	Lower	Upper	SD	Minimum	Maximum
Creativity	GPT-4	247	0.632	0.622	0.642	0.0795	0.477	0.932
	Human	246	0.738	0.728	0.749	0.0827	0.530	1.068
Flexibility	GPT-4	247	0.934	0.931	0.937	0.0231	0.851	0.992
	Human	246	0.899	0.891	0.906	0.0626	0.638	1.038

Note. The subjective ratings of creativity were made on 3-point scale (0 = *not at all*, 2 = *very*). Theoretically, the semantic distance may vary between 0 and 2, with higher scores indicating higher distance.
[a]*N* for AI refers to the number of test sessions.

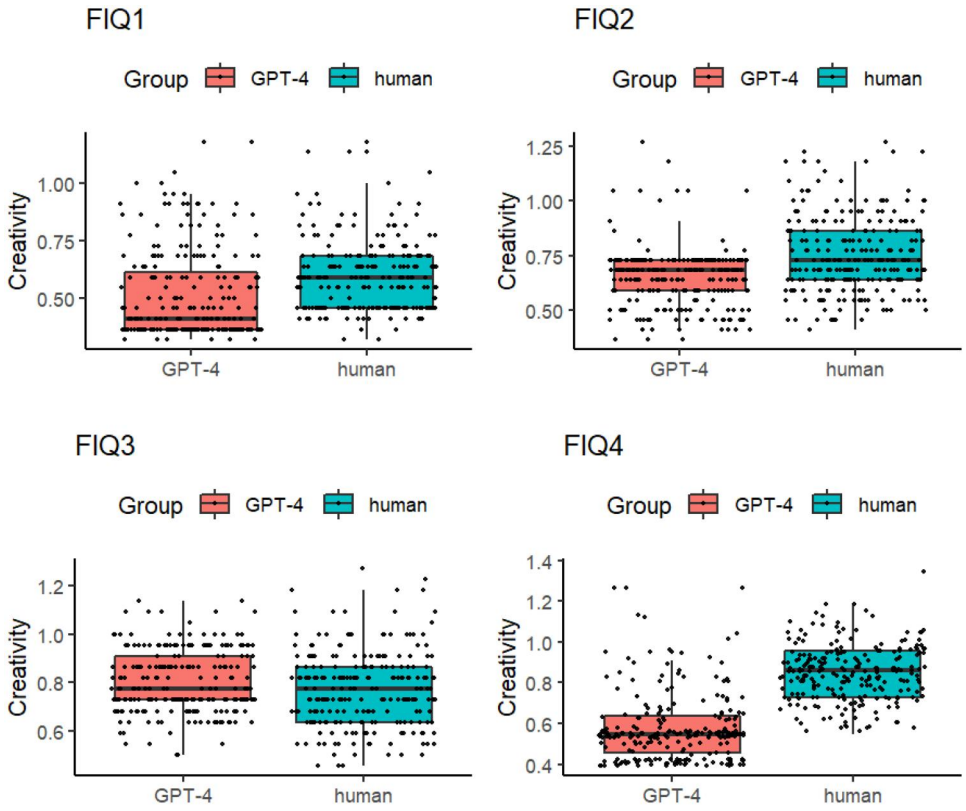

Figure 3. GPT-4's and humans' creativity scores in response to the four figures.

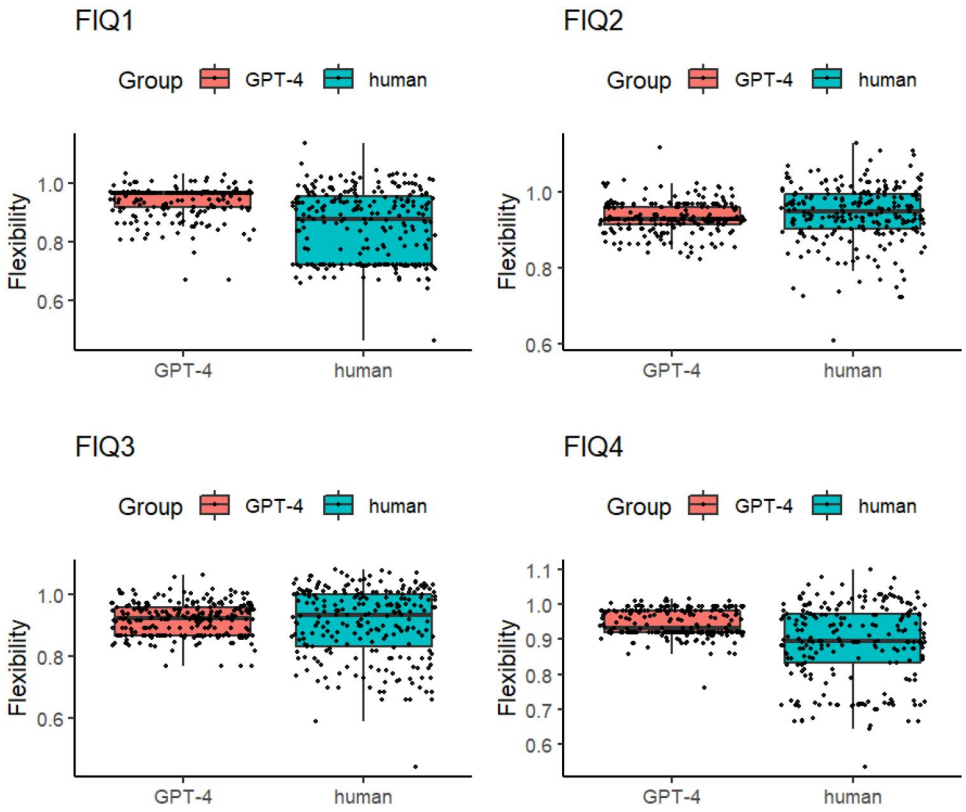

Figure 4. GPT-4's and humans' flexibility scores in response to the four figures.

FIQ4, B = 0.056, SE = 0.007, 95% CI [0.042, 0.070], $t(1885) = 7.60$, $p < .0001$. However, no statistically significant differences were detected in response to FIQ2, B = −0.011, SE = 0.007, 95% CI [−0.025, 0.004], $t(1886) = -1.41$, $p = .158$, and to FIQ3, B = 0.010, SE = 0.007, 95% CI [−0.005, 0.025], $t(1886) = 1.36$, $p = .175$.

4. Discussion

The present study explored the comparative creative capacities of artificial intelligence (AI) and humans using the Figural Interpretation Quest (FIQ), a newly developed multimodal task for testing creative behavior. Our findings contribute to the ongoing discourse on the role and capabilities of AI in creative domains, a topic that has garnered significant attention during the past year (Oksanen et al., 2023).

Our results showed that ChatGPT-4 showcased a higher average level of flexibility compared to humans in generating diverse interpretations of the nonsense FIQ figures. This result, as shown in Figure 2(B), is in line with what was predicted in H1. However (see Figure 2(A)), human participants scored on average higher in creativity, when the creativity of the responses were subjectively assessed by other humans. This result goes against the prediction of H1, for which we expected AI to generally outperform human average creativity scores independently on the metric used. Best humans still outperformed AI, both in subjectively perceived creativity and in flexibility (see the highest scores in Figure 2(A and B)), in line with H2. The correlation analyses, investigating the relationship between creativity and flexibility scores, showed a weak correlation in human participants ($r = .175$, $p = .013$). Furthermore, no statistically significant correlation between the measures was found for the AI outputs ($r = .071$, $p = .498$). These results refute the exploratory H3 and suggest a more complex relationship between divergent thinking and perceived creativity than initially hypothesized.

Compared to previous findings using text-based creativity assessments such as the AUT (Haase & Hanel, 2023; Koivisto & Grassini, 2023) and the verbal form of the Torrance Tests of Creative Thinking (Guzik et al., 2023), and in which AI's creative productions were on average preferred against human ones, our study indicates a different pattern. Our results showed that when subjectively evaluating creative outputs, raters displayed a preference for human-generated content over AI-produced work. These results may be due to the tasks used in this study to test creativity (FIQ), which emphasize multimodal cognition and might underscore a human advantage in processing visual stimuli. This could be related to the role of imagery in certain creative tasks. We propose that creative imagery is crucial in performing the FIQ tasks, and therefore, our results may help understanding the difference between AI and human creativity. Unlike AI, which produces responses based on statistical data patterns, human creativity draws from a diverse range of subjective experiences and sensory-driven cognitive processes. On the other hand, Chatbots excel in verbal-type tasks which only rely on symbol manipulation-type of processes and do not require a sensory-based cognition that is impossible for machines that do not possess sensory experiences, and that so far are unable to fully simulate such type of human cognitive processes.

The data presented in this study illustrate a complex picture of AI's creative abilities as compared to human creativity. While the AI Chatbot ChatGPT-4 may show superiority in average performance across some dimensions of creative behavior such as flexibility, our findings suggest that this may not equate to generally surpassing human creativity. Understanding the type of creativity (and in general cognitive abilities), where AI excels compared to humans is important for establishing the possible scope and limitations of AI in practical contexts. Such clarification may contribute to the public debate about the role of AI in society. Specifically, our findings indicate that in complex interpretative multimodal tasks as the proposed FIQ, the superior performances of the best human participants still surpass AI capabilities. This is in line with previous studies that have used other creative tasks (Guzik et al., 2023; Haase & Hanel, 2023; Koivisto & Grassini, 2023). Our results suggest that while AI can generate creative outputs that are remarkable on average, it has not yet reached the highest level of human creativity.

Creativity is commonly defined as the capacity to generate ideas that are somewhat original (Runco & Jaeger, 2012). This explanation focuses on the outcomes of creativity, judging an individual's creative skills based on the inventive results produced, rather than detailing the internal mechanisms that lead to these creative thoughts. Hence, the empirical evidence presented in the present study corroborates previous studies (Guzik et al., 2023; Koivisto & Grassini, 2023) that AI's capability to generate creative outputs that has matched or even surpassed the average human level in certain type of tasks. The present study suggests that creative tasks in which flexibility is required are these where chatbots excel best. The inherent strengths of current AI technologies, including their expansive memory and rapid access to extensive text databases, give AI an advance against humans, when the performance in the creative task requires flexibility. These capabilities enable AI systems to excel in associative aspects of divergent thinking, a key component of creativity that involves generating semantically diverse ideas. Overall, as in the findings in the arts (Arriagada, 2020), our evidence proposes that the ability to produce creative ideas might not be exclusively inherent to conscious human beings.

Furthermore, the finding that human raters found human-generated interpretations of the figures to be more creative than those provided by AI could be attributed to the larger degree of variation of human responses (although a particular person's responses to a specific figure were not necessarily very flexible, that is, semantically distant from each other), and not to objective differences in the quality of these responses when it comes to creativity. The larger degree of variation of human responses compared to AI ones can be observed from the larger ranges of the boxplots in Figure 2(B). This disparity in creativity richness between humans and machines can be rooted in the very nature of human creativity itself and in the way it may have evolved. Evolution promotes the diversification of traits, as a broader range of characteristics can improve survival chances. Consequently, it is unsurprising that human creativity tends to exhibit a more diverse range of ideas and concepts. In the present type of tasks, raters may find the less common ideas more creative, influenced by a "rarity effect."

This effect may have led to a preference for types of responses that are more semantically different from each other when encountered in the list of interpretations provided to the raters (independently on the quality of the interpretations or the semantic distance between response couples).

In contrast, AI chatbot models are engineered to produce statistically optimal responses. These responses, while varied to some extent, tend to show a higher degree of similarity when a similar prompt is given to the AI chatbot. As a result, AI tends to generate responses that are similar (between prompts) when prompted with the same task, especially for tasks requiring brief answers, such as those with a one or two-word limit (in fact the response range of AI was roughly half of those provided by humans). This leads to a narrower range of diversity in AI responses compared to those generated by humans, that are instead influenced by a complex mix of evolutionary and cognitive factors that drive a richer diversity in creative interpretations.

The novelty of the present study compared to the previous ones is that it used newly developed creativity tasks, that therefore is less likely for the chatbot to have "seen" optimized responses during in its training data. In fact, according to ChatGPT-4, the training material, during the experiment testing, was up to September 2021, while the FIQ task was published in 2022 (Erwin et al., 2022; Koutstaal et al., 2022). Furthermore, this is the first study we are aware of in scientific literature, that challenges a chatbot in a multimodal creativity task that does not just involve words but also interpretations of images.

4.1. Limitations and future research

Several limitations should be considered when interpreting the results of the present study. Firstly, our analysis was confined to only four out of the numerous potential figures presented in the FIQ. While the chatbot asserts no training data access post-September 2021, validating this claim is beyond our possibility as OpenAI – the company behind ChatGPT-4 – does not explicitly state this information. Thus, it's conceivable that it had some exposure to FIQ-related information. However, considering the FIQ's recent development and the so far little published material on it, any such exposure in the chatbot's training data is unlikely to have significantly distorted its performance.

Additionally, it's important to note the variability of our results across different FIQ figures. While human interpretations were generally perceived as more creative for Figures 1, 2, and 4, the same human raters assigned higher creativity scores to AI-generated interpretations for Figure 3. This variation underscores that even within the same task type, factors such as randomness, the nature of the AI's training data, or unknown aspects of the AI's architecture may lead to instances where AI outperforms the average human. This is particularly notable in the context of subjective assessments of creativity made by human raters. This finding highlights the complexity and multifaceted nature of creativity assessment.

Furthermore, the FIQ task has been proposed as a measure of divergent thinking (Erwin et al., 2022). However, divergent thinking should not be considered "creativity" but only one of its components. Future studies should assess how findings regarding divergent thinking abilities in AI may expand to other domains of creativity.

Human participants who didn't complete the task of providing two creative interpretations for at least one figure were excluded. This exclusion assumed that incomplete responses stemmed from distraction or low attention. Yet, it is plausible that some participants' non-completion was due to a complete lack of creative outputs, suggesting they might have been included in the analysis with a minimum score.

In our study, we controlled for both fluency and elaboration. We hypothesized that AI, particularly chatbots, might possess an inherent advantage in these domains due to their capacity for quick and coherent text generation. Thus, fluency could undermine the assessment of creativity (Dygert & Jarosz, 2020). To mitigate this, we intentionally structured the task to limit fluency, requiring only two interpretations for each FIQ figure. Additionally, we customized the task's design, especially in terms of the allowed response length for AI figure interpretation, to align with human response capabilities. This adjustment was crucial because elaboration (i.e. the length of responses) could be a confounding variable in the subjective measurement of creativity (Dippo & Kudrowitz, 2015). Without this control, the AI's ability to quickly generate text may have had given the raters an impression of higher elaboration in the cases of AI's responses, thus giving for AI an unfair advantage over human participants. However, this aspect of the experimental design may have inadvertently disadvantaged the chatbot by tailoring the task parameters more closely to human rather than AI capabilities.

Future research should seek to build on the present findings by evaluating AI and human performance in a wider variety of tasks designed to evaluate creativity and updating our results considering future and more advanced AI technologies. It is crucial to assess AI's performance across a spectrum of multimodal and sensory-based creative tasks (e.g. using sounds), which are expected to become feasible with forthcoming advancements in AI capabilities. Such investigations are crucial for understanding the variety of AI's strengths and limitations when it comes to creativity. Moreover, it is essential to investigate the ways in which humans can use the types of creativity where AI shows advantage, to augment human creative processes. This line of investigation holds the promise of encouraging a synergistic interaction between human and machine cognitive abilities, with the potential of improving human performances.

5. Conclusion

In conclusion, this study reported an assessment of the creative capabilities of artificial intelligence (AI), with a focus on ChatGPT-4, in comparison to human creativity. Utilizing the newly developed Figural Interpretation Quest (FIQ) as a benchmark, this research stands out as the first to employ a multimodal task combining visual and textual elements. ChatGPT-4 exhibited significant strengths in the creativity feature of flexibility, demonstrating an ability to generate a

variety of creative ideas with distinct meanings superior to the average human. However, it did not consistently beat human creativity when subjectively perceived creativity was evaluated by human raters. Notably, the most exceptional human responses demonstrated superior performance in terms of subjectively assessed creativity and flexibility, compared to AI outputs. This finding underscores the complex landscape of creative expression, where AI shows considerable promise, even beating human being in some specific areas or tasks, while human creativity, despite the remarkable development of AI, being able to outperform AI in other aspects of creative behavior. We proposed that differences in cognitive architectures between humans and machines may be the culprit of such differences.

On a practical level, our findings indicate that current AI technologies cannot entirely replace human creativity but rather serve to complement and enhance it, especially in areas where chatbots like ChatGPT-4 excel. By leveraging on AI's distinctive strengths, these technologies can be integrated into human creative activities and works, creating new possibilities for human–AI collaboration across various creative domains. Future research should explore whether this balance remains with the introduction of more advanced AI models and those based on different architectures.

Note

1. The figure used in the present study were the ones with the codes A1_cat2508, C1_cat2008, C3_cat1105, D1_cat1807 from the FIQ dataset as shared with us by the corresponding author of Erwin et al. (2022).

Acknowledgements

The authors thank Maria Ulvang, Preben Bjarkøy Jensen, and Markus Malkenes Sørtun, Master students at the University of Bergen, for their valuable feedback during the testing of the subjective creativity rating task.

Author contributions

Conceptualization, S.G., and M.K., methodology, S.G. and M.K.; formal analysis, M.K. and S.G.; investigation, S.G., and M.K.; writing – original draft, S.G., and M.K.; writing – review and editing, S.G. and M.K. Both authors have read and agreed on the submitted version of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Simone Grassini http://orcid.org/0000-0002-4189-7585
Mika Koivisto http://orcid.org/0000-0002-5227-0091

Data availability

Data supporting the findings of the present article can be found following this url: https://osf.io/download/ap3ur/?view_only=214c1de69fb146e099968e75382b0515

Code availability

The used R scripts for data analysis can be found following this url: https://osf.io/download/9pj4s/?view_only=214c1de69fb146e099968e75382b0515

References

Anantrasirichai, N., & Bull, D. (2021). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review*, *55*(1), 589–656. https://doi.org/10.1007/s10462-021-10039-7

Antonietti, A. (1991). Chapter 15 Why does mental visualization facilitate problem-solving? In *Mental images in human cognition* (pp. 211–227). Elsevier.

Antonietti, A., & Colombo, B. (1997). The spontaneous occurrence of mental visualization in thinking. *Imagination, Cognition and Personality*, *16*(4), 415–428. https://doi.org/10.2190/DYHQ-XMA7-YQW1-NT14

Antonietti, A., & Colombo, B. (2011). Mental imagery as a strategy to enhance creativity in children. *Imagination, Cognition and Personality*, *31*(1), 63–77. https://doi.org/10.2190/IC.31.1-2.g

Antonietti, A., Colombo, B., & Pizzingrilli, P. (2020). The mechanisms of creative thinking. In A. Antonietti, P. Pizzingrilli, C. Valenti (Eds.), *Enhancing creativity through story-telling. Palgrave studies in creativity and culture* (pp. 1–14). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-63013-3_1

Arriagada, L. (2020). CG-Art: Demystifying the anthropocentric bias of artistic creativity. *Connection Science*, *32*(4), 398–405. https://doi.org/10.1080/09540091.2020.1741514

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*. https://doi.org/10.48550/arXiv.1506.04967

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, *53*(2), 757–780. https://doi.org/10.3758/s13428-020-01453-w

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, *42*(7), 1186–1197. https://doi.org/10.3758/s13421-014-0428-8

Binder, W. (2020). AlphaGo's deep play: Technological breakthrough as social drama. In *The cultural life of machine learning* (pp. 167–195). Springer International Publishing.

Carnovalini, F., & Rodà, A. (2020). Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence*, *3*, 14. https://doi.org/10.3389/frai.2020.00014

Chamberlain, R., Mullin, C., Scheerlinck, B., & 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. https://doi.org/10.1037/aca0000136

Chi, M. T. H. (1997). Creativity: Shifting across ontological categories flexibly. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 209–234). American Psychological Association.

Cho, S. H., Nijenhuis, J. T., Van Vianen, A. E. M., Kim, H. B., & Lee, K. H. (2010). The Relationship between diverse components of intelligence and creativity. *The Journal of Creative Behavior*, *44*(2), 125–137. https://doi.org/10.1002/j.2162-6057.2010.tb01329.x

Clément, E. (2022). Successful solution discovery and cognitive flexibility. In *Cognitive flexibility* (pp. 113–142). Wiley.

Dippo, C., & Kudrowitz, B. (2015). The effects of elaboration in creativity tests as it pertains to overall scores and how it might prevent a person from thinking of creative ideas during the early stages of brainstorming and idea generation. *27th International Conference on Design Theory and Methodology*, *7*. https://doi.org/10.1115/DETC2015-46789

Duca Iliescu, D. M. (2020). The impact of artificial intelligence on the chess world. *JMIR Serious Games*, *8*(4), e24049-e24049. https://doi.org/10.2196/24049

Dygert, S. K. C., & Jarosz, A. F. (2020). Individual differences in creative cognition. *Journal of Experimental Psychology. General*, *149*(7), 1249–1274. https://doi.org/10.1037/xge0000713

Erwin, A. K., Tran, K., & Koutstaal, W. (2022). Evaluating the predictive validity of four divergent thinking tasks for the originality of design product ideation. *PloS One*, *17*(3), e0265116-e0265116. https://doi.org/10.1371/journal.pone.0265116

Finke, R. A. (1996). Imagery, creativity, and emergent structure. *Consciousness and Cognition*, *5*(3), 381–393. https://doi.org/10.1006/ccog.1996.0024

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

Georgsdottir, A. S., & Getz, I. (2004). How flexibility facilitates innovation and ways to manage it in organizations. *Creativity and Innovation Management*, *13*(3), 166–175. https://doi.org/10.1111/j.0963-1690.2004.00306.x

González, M. A., Campos, A., & Pérez, M. J. (1997). Mental imagery and creative thinking. *The Journal of Psychology*, *131*(4), 357–364. https://doi.org/10.1080/00223989709603521

Grajzel, K., Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2023). Measuring flexibility: A text-mining approach. *Frontiers in Psychology*, *13*, 1093343. https://doi.org/10.3389/fpsyg.2022.1093343

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, *13*(7), 692. https://doi.org/10.3390/educsci13070692

Guilford, J. P. (1967). Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, *1*(1), 3–14. https://doi.org/10.1002/j.2162-6057.1967.tb00002.x

Guzik, E. E., Byrge, C., & Gilde, C. (2023). The originality of machines: AI takes the Torrance Test. *Journal of Creativity*, *33*(3), 100065. https://doi.org/10.1016/j.yjoc.2023.100065

Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, *33*(3), 100066. https://doi.org/10.1016/j.yjoc.2023.100066

Hacker, P. (2021). A legal framework for AI training data—From first principles to the Artificial Intelligence Act. *Law, Innovation and Technology*, *13*(2), 257–301. https://doi.org/10.1080/17579961.2021.1977219

Hitsuwari, J., Ueda, Y., Yun, W., & 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. https://doi.org/10.1016/j.chb.2022.107502

Jaušovec, N. (1991). Flexible strategy use: A characteristic of gifted problem solving. *Creativity Research Journal*, *4*(4), 349–366. https://doi.org/10.1080/10400419109534411

Jaušovec, N. (1994). Can giftedness be taught? *Roeper Review*, *16*(3), 210–214. https://doi.org/10.1080/02783199409553576

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences*, *23*(4), 271–274. https://doi.org/10.1016/j.tics.2019.01.007

Kim, D. H. (2023). A study on the job replacement impact of ChatGPT and education method. *European Journal of Science, Innovation and Technology*, *3*(4), 105–122. https://ejsit-journal.com/index.php/ejsit/article/view/245

Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, *13*(1), 13601–13601. https://doi.org/10.1038/s41598-023-40858-3

Kosslyn, S. M. (1980). *Image and mind*. Harvard University Press.

Koutstaal, W., Kedrick, K., & Gonzalez-Brito, J. (2022). Capturing, clarifying, and consolidating the curiosity-creativity connection. *Scientific Reports*, *12*(1), 15300–15300. https://doi.org/10.1038/s41598-022-19694-4

Koutstaal, W., Reddy, C., Jackson, E. M., Prince, S., Cendan, D. L., & Schacter, D. L. (2003). False recognition of abstract versus common objects in older and younger adults: Testing the semantic categorization account. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(4), 499–510. https://doi.org/10.1037/0278-7393.29.4.499

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, E. (2011). Digital originality. Vanderbilt Journal of Entertainment and Technology Law, *14*, 919. https://scholarship.law.vanderbilt.edu/jetlaw/vol14/iss4/5

Mazzone, M., & Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. *Arts*, *8*(1), 26. https://doi.org/10.3390/arts8010026

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232. https://doi.org/10.1037/h0048850

Mehta, R., & Dahl, D. W. (2018). Creativity: Past, present, and future. *Consumer Psychology Review*, *2*(1), 30–49. https://doi.org/10.1002/arcp.1044

Millet, K., Buehler, F., Du, G., & Kokkoris, M. D. (2023). Defending humankind: Anthropocentric bias in the appreciation of AI art. *Computers in Human Behavior*, *143*, 107707. https://doi.org/10.1016/j.chb.2023.107707

Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y., & Savela, N. (2023). Artificial intelligence in fine arts: A systematic review of empirical research. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100004. https://doi.org/10.1016/j.chbah.2023.100004

Ovando-Tellez, M., Kenett, Y. N., Benedek, M., Bernard, M., Belo, J., Beranger, B., Bieth, T., & Volle, E. (2022). Brain connectivity-based prediction of real-life creativity is mediated by semantic memory structure. *Science Advances*, *8*(5), eabl4294-eabl4294. https://doi.org/10.1126/sciadv.abl4294

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, *3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Purves, D. (2019). Opinion: What does AI's success playing complex board games tell brain scientists? *Proceedings of the National Academy of Sciences of the United States of America*, *116*(30), 14785–14787. https://doi.org/10.1073/pnas.1909565116

R Core Team. (2018). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riquelme, H. (2002). Can people creative in imagery interpret ambiguous figures faster than people less creative in imagery? *The Journal of Creative Behavior*, *36*(2), 105–116. https://doi.org/10.1002/j.2162-6057.2002.tb01059.x

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. https://doi.org/10.1080/10400419.2012.650092

Runco, M. A., & Okuda, S. M. (1991). The instructional enhancement of the flexibility and originality scores of divergent thinking tests. *Applied Cognitive Psychology*, *5*(5), 435–441. https://doi.org/10.1002/acp.2350050505

Samuelson, P. (2023). Generative AI meets copyright. *Science (New York, N.Y.)*, *381*(6654), 158–161. https://doi.org/10.1126/science.adi0656

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Spector, J. M., & Ma, S. (2019). Inquiry and critical thinking skills for the next generation: From artificial intelligence back to human intelligence. *Smart Learning Environments*, *6*(1), 1–11. https://doi.org/10.1186/s40561-019-0088-z

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/brm.42.4.1096

Stoet, G. (2016). PsyToolkit. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Wilkenfeld, Y. (2019). Can chess survive artificial intelligence? *The New Atlantis*, *58*, 37–45. https://www.jstor.org/stable/26609113

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, *14*, 1181712. https://doi.org/10.3389/fpsyg.2023.1181712

About the authors

Simone Grassini works as an associate professor at the Department of Psychosocial Science, University of Bergen (Norway), and he specializes in psychology and cognitive science. One of his research lines during the past few years has been studying the interaction between human cognition and emerging technologies, including VR and AI.

Mika Koivisto works as a senior lecturer at the Department of Psychology, University of Turku (Finland). His major research line during the last 20 years has focused on the neural correlates of consciousness and on the distinction between conscious and unconscious processing.