Explorations in human vs. generative AI creative performances: A study on human-AI creative potential

Mia Magdalena Bangerl*
mia.bangerl@tugraz.at
Graz University of Technology &
Know-Center GmbH
Graz, Austria

Katharina Stefan
katharina.stefan@edu.uni-graz.at
University of Graz
Graz, Austria

Viktoria Pammer-Schindler
viktoria.pammer-schindler@tugraz.at
Graz University of Technology &
Know-Center GmbH
Graz, Austria

ABSTRACT

Advances in quality and range of generative AI have opened up new possibilities for AI-supported work and human-AI collaboration. Now, researchers are challenged to investigate how, where, and to whom AI can contribute meaningfully. In this paper, we present a study on human versus AI creative performance in the Alternate Uses Test (AUT) and discuss the implications of our results for human-AI collaboration. We analyze how different text-generative AI chatbots compare to human dyads in the AUT regarding creative fluency, originality, flexibility, and elaboration. Our results reveal high ranges in performance within both the human dyad group and the AI chatbot group. Further, humans excel in original and flexible ideation, while AI better elaborates and details responses. Therefore, collaborative creative performance in human-AI teams could benefit from these different but complementary skills. In future work, we will test this assumption and explore the social dynamics of human-AI collaboration to find ways of trustworthy and reliable human-AI collaboration.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative interaction**; *Empirical studies in HCI*.

KEYWORDS

generative AI, collaborative creativity, computational creativity, artificial intelligence, divergent thinking

ACM Reference Format:
Mia Magdalena Bangerl, Katharina Stefan, and Viktoria Pammer-Schindler. 2024. Explorations in human vs. generative AI creative performances: A study on human-AI creative potential. In *TREW 2024: Trust and Reliance in Evolving Human-AI Workflows, at CHI 2024, May 11, 2024.* ACM, New York, NY, USA, 8 pages.

1 INTRODUCTION

Since the introduction of AI-Chatbot ChatGPT-3 [26] in late 2022, and the vast progress in AI development ever since, generative AI technology has been a central point of societal, political, and scientific discussion. A substantial aspect of this discourse is how AI can, should, and will integrate into the dynamics of modern society, e.g. different fields of work. As AI performance tends to vary greatly between different tasks and domains [10], research is challenged to explore how and for whom AI can make a meaningful contribution.

In our work, we are interested in the potential of generative AI for different creative tasks. Existing investigations of AI's creative potential include studies on divergent thinking [23, 32], creative writing [11, 16, 18, 36], visual art and design[3, 8], music making [24], and dance [35]. Within the realm of creative tasks, we are specifically interested in divergent thinking. Our overall aim is to investigate how the creative performance of one, or multiple AIs compares to human performance, how AI creativity is different from human creativity, and if human-AI collaboration can enhance creative performance, and potentially even inspire new forms and processes of creativity.

In this paper, we discuss preliminary work that compares the task performance of 20 human dyads (pairs) in the Alternate Uses Test (AUT) by Guilford [14] with the performance of 10 text-generative AI Chatbots. In doing so, we want to find out how AI creative performance differs from human creative performance and explore areas of potential for human-AI collaboration. The results of this study will contribute to a better understanding of how AI can be used effectively and sustainably to support and enhance human creative work. Our work improves understanding of strengths and shortcomings of both human and AIs' creative performance strategies and outcomes and thus also offers insights into the potential of human-AI creative collaboration in the future.

In the following Section 2, we will briefly revisit the related work on human-AI creative collaboration and human vs. AI creative performance. We introduce our research question in Section 3, and describe our study and analyses in Section 4. In our results Section 5, we present the outcome of the study on human vs. AI creative performance, and then finally discuss our results and implications for future work in Section 6.

2 RELATED WORK

2.1 Creativity and Creative Collaboration

Creativity or creative performance, meaning the development and creation of something new, purposeful, and meaningful [27], has long been attributed almost entirely to humans. But in contrast to this belief, computational creativity has evolved from an inferior creative performance to one rivaling and, in some instances, even

CHI EA'24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 978-1-4503-XXXX-X/18/06

surpassing human performance in comparative creativity studies [18, 23].

Creative collaboration, also called distributed creativity, [30] refers to situations in which a creative product emerges from the collaboration that is more than the sum of each collaborator's contribution. Collaboration has to be distinguished from mere cooperation, which refers to situations where the creative product does not go beyond the sum of the collaborator's contributions, the overall labor is just divided between contributors, and parts can be done separately and asynchronously [7, 27].

2.2 Computational Creativity

Following Maher et al. [25], we refer to computational creativity as creativity that is created and evaluated in computational systems and includes both a creative process and product. Because different computational systems can assume different roles and functions in a creative process, Davis et al. [7] distinguish between three types of computational creativity: 1) Support tools, which can be used to support and extend human-made creative work, e.g., image editing software; 2) Generative systems, which can generate creative products largely on their own, e.g., generative AI; and 3) Computer colleagues, that can collaborate synchronously with human collaborators, to achieve a joint creative product, e.g., the co-creative movement AI Lumin [35]. In this work, we investigate the creative performance of generative systems. Our results will also inform future explorations regarding the potential of generative AI as a computer colleague for creative work.

Today, a multitude of computational systems for human-computer collaborative creativity are available, covering a broad range of creative disciplines, including e.g. creative writing, visual arts and designing, acting, dancing, speaking, and music making [27]. Some systems featured in the scientific literature are the Drawing Apprentice for collaborative human-computer digital drawing [9], the music editing AI Cococo [24], and the collaborative human-AI story-writing tool Wordcraft [36]. Furthermore, Haase et al. [17] have published on the use of image-generation AI DALL-E-2 for human-AI creative inspiration, and Winston & Magerko [35] have studied turn-taking based on the movement and dance-based AI Lumin. These works investigate the promises and potentials of AI for creative work and human-AI creative collaboration, e.g., by stimulating and diversifying creativity in a collaborative setting [21, 25].

Nonetheless, there are several concerns regarding the use of and collaboration with AI, both in general and for creative work. One area of worry is the long-term influence of AI use on humans. Both Sætra [33] and Castro et al. [6] discuss a potential reduction in abilities that are taken over by AI (such as writing skills). More broadly, this could result in a homogenization of texts, art, and speech, and consequential cultural impoverishment. Furthermore, AI use in human-to-human communication and collaboration might harbor mistrust and paranoia, thus harming social dynamics and collaboration quality [19]. Finally, AI represents only a very narrow percentage of humans and carries biases accordingly, as found by Atari et al. [4] who conclude that ChatGPT answers represent predominately Western, educated, industrialized, rich, and democratic populations.

A debated question is also how, and whether, AI can be creative. Human creativity can be inspired by experience, feelings, external environments, material objects, and interaction with others [7, 22]. AI creative work, however, is always based on existing data, thus Kirkpatrick [22] concludes that nothing produced by AI is truly original – AI mimics creativity, rather than self-creating.

2.3 Performance in computational creativity

In this study, we are concerned with the performance of modern text-generative AI in creativity. This interest follows up on existing work that has investigated human vs. AI performance in different creative tasks:

A study by Weingarten et al. published in 2020 [34] found that the performance of human experts in logo design was rated higher than that of AI. In 2022, Stevenson et al. [32] also evaluated AI answers to the Alternate Uses Test as less original and overall less creative than those answers given by humans. In the same year, the writing quality, beauty, and interestingness of AI-generated text were also ranked worse than human writing in a study by Gunser et al. [16]. Then, in 2023, Hitsuwari et al. [18] found that human and AI-written haikus were rated similarly regarding beauty, and were also hard to distinguish by raters. Also, Kovisto & Grassini [23] found that AI now performed better than most humans in the Alternate Uses Test, though top-performing humans still ranked higher. However, they also state that it is unclear how AI comes up with answers, whether they are created originally, or simply copied from existing data.

Overall, it seems that the most up-to-date studies observe performances by generative AI that have reached the human levels in some instances, although human experts still match and sometimes outshine the AI. Despite these results, our understanding of AI creativity is still blurry and incomplete. Though AI can generate creative ideas and products, it is unclear how original its work is, how flexible it is in the long run, and under which conditions it becomes repetitive and monotone. Also, regarding human-AI collaboration, interestingly not many studies focus on collaboration in the sense of computer colleagues [7], and rather study AI in either assisting or tutoring roles.

For these reasons, as well as the rapid progress of AI development, we are interested in continuing research on AI vs. human creative performance, as a basis for exploring the dynamics and creative outcome of human-AI collaboration in the future.

3 RESEARCH QUESTION

In the study presented in this paper, we ask the following research question:

RQ: How do different text-generative AI chatbots compare to human dyads in the Alternate Uses Test (AUT) regarding fluency, originality, flexibility, and elaboration?

We will answer our research question by comparing 20 human dyads (pairs) and 10 AI chatbots[1] in the dimensions fluency (the

[1]Note that we compare human dyads with singular AI chatbots instead of comparing e.g. individual humans and chatbots, or two pairs. This study design was chosen in anticipation of future studies on the potentials (and risks) of AI use in human-to-human collaboration, as the collaborative creative effort is central to this research endeavor. Further, it is questionable whether true collaboration can be achieved between two AI chatbots in the first place, given our definition of collaboration discussed in Section 2.1.

number of alternate uses found for a designated object), originality (the rarity of the found categories of use compared to all other dyads/bots), flexibility (the range of categories of use found by one dyad/bot) and elaboration (the level of detail in explaining the use). This study follows up on the above-described work on computational creative performance in Section 2.3, and adds a comparison between multiple systems as well as a discussion on the performance of the human cohort in relationship to each generative AI system.

4 MATERIALS AND METHODS

4.1 Alternate Uses Test (AUT)

The Alternate Uses Test [14] is an established task for measuring creative potential [28]. To complete the task, participants are asked to name as many non-typical uses for a specific object (e.g. a tennis ball) as they can think of, producing a list of alternate uses, which can then be assessed to measure the creative potential. The AUT has been selected for the study due to its focus on divergent and creative thinking. Furthermore, the AUT can be done by both human and AI test subjects, as it does not rely on material objects or physical movement. It has also previously been used to study human vs. AI creative performance [23, 32], thus, results from these previous studies can be used as comparisons.

4.2 Experimental procedure

Human dyads: The uses generated by human dyads were taken from a previous study on synchronization in face-to-face vs. video conferencing-mediated collaboration conducted in Oct-Nov 2023, in which dyads worked on four collaborative tasks, one of them being the AUT. For this research endeavor, we only extracted the uses generated by the face-to-face group, which consisted of 20 dyads. In the task, two participants sat next to each other, in front of one computer display, where they received written online task instructions. They had a time window of 3 minutes per object and consecutively worked on three AUT objects: fork, balloon, and key. The dyads did not have to write down their found uses but verbally brainstormed and discussed. They received instructions and communicated in the native language, German. The generated uses were later transcribed for analysis.

AI chatbots: The uses generated by the AI chatbots were collected separately for this study, between January 19th and February 5th, 2024. We gave the chatbots an adapted translated version of the instructions for the dyads. Adaptations were that the bots performed the task individually, in writing, and in English (the default AI language). Because the AI chatbots gave different numbers of uses after the initial prompt, we asked them repeatedly to find more alternate uses, until they reached or surpassed the *median* number of uses generated by the human dyads ($md_{\text{fork}} = 13$, $md_{\text{balloon}} = 12.5$, $md_{\text{key}} = 12$), refused to answer the prompt, or exactly repeated their previously generated uses. We always used the same wording for the initial prompt, the repeat prompt, and the transition toward the next object. Additionally, we ensured not to ask the chatbots anything else, before the task execution, to ensure an unbiased start.

4.3 Participants

Human dyads: The 40 human participants were students from an Austrian university (mostly psychology students), who gave their informed, written consent before the experiment and were rewarded with a study participation certificate[2]. The study was approved by the ethics committee of the responsible university. The participants were paired into dyads for the study, based on their results to the AIST-R general interest-structure test [5], which can be used to calculate the interest profiles of individuals [20]. Participants were matched to foster good collaborative conditions, such as shared interest, empathy, and sympathy, and results from the post-experiment questionnaire confirmed high partner sympathy ($mean = 4.18$ on a scale from 1 = lowest to 5 = highest partner sympathy). We ensured that participants did not know each other, before testing. The gender distribution in the sample was 62.55% female, and 37.5% male, and the participant's ages were between 18 and 32 years ($mean = 22.28$, $md = 22$, $sd = 3.20$). Most participants' highest formal education was the completion of Austrian secondary school (62.50%), 19.5% had completed an apprenticeship or compulsory school, and 17.5% already had an academic degree. Questionnaire answers also show that the human group was moderately interested in the AUT, as the *mean* interest rate on a scale from 1 (very low interest) to 5 (very high interest) was 3.80 ($md = 4$, $sd = 0.61$).

AI chatbots: The selection of the 10 text-generative AI chatbots, displayed in Table 1, was based on three main criteria. Firstly, the chatbot had to be publicly available and usable free of charge, so we excluded all pay-to-use AI models, though free trials or demos were included. Secondly, the chatbot had to be able to talk about any topic if asked, and should not be designed for specific use cases only, so this excluded e.g. health, care, and therapeutic support chatbots, like Sophia, a chatbot for supporting victims of domestic violence [1], educational and tutoring systems, like the GitHub Copilot for coding assistance [13] and customer support/marketing chatbots, like Drift [12]. Thirdly, the chatbot had to be proficient in English, to ensure language homogeneity within the AI chatbot sample.

4.4 Data collection

The collected data consists of the alternate uses found by the human dyads and the AI chatbots, for the three objects: fork, balloon, and key. In total, the 20 dyads generated 783 valid uses, and the 10 chatbots generated 395 valid uses altogether. Based on this data, we will answer our research question regarding potential differences in the performance of human dyads vs. AI chatbots in the AUT. For the human dyads only, we also collected demographic data and task interest in questionnaire form, among other information that is not relevant to the research question of this paper.

4.5 Data analysis

The analysis of data in this study consisted of an initial scoring of the alternate uses generated per AUT object by each human dyad/AI chatbot, which were then summed up to provide one score per dyad/AI for each of the four AUT dimension, as well as the

[2]The psychology students were required by their university to participate in several scientific studies as part of the curriculum. The study participation certificate serves as recognized proof of such participation.

Table 1: List of AI Chatbots Selected for the Experiment

AI chatbot	Owner	Testdate	Note
Microsoft Copilot	Microsoft	19.01.2024	We selected the mode "creative".
ChatGPT-3.5	OpenAI	19.01.2024	
Bard	Google	19.01.2024	Bard has been renamed "Gemini".
Llama2	Meta	19.01.2024	
Pi	Inflection AI	19.01.2024	
Character Assistant	Character. AI	19.01.2024	
Perplexity AI	Perplexity	26.01.2024	
Copy AI	CopyAI	02.02.2024	
Claude	Anthropic	05.02.2024	We used Claude-instant via Poe.com.
Vicuna	LMSYS Org	05.02.2024	We used Vicuna-13B.

overall score sum of all dimensions. After calculating these scores, we used measures of central tendency and analyzed statistical distributions to compare the two experimental groups. Furthermore, we calculated Mann-Whitney U tests (due to non-parametric data distribution) to detect significant group differences.

For scoring the generated uses regarding fluency, originality, flexibility, and elaboration, we adopted our detailed method of data analysis, including the point-wise weighting of the categories, from Alhashim et al. [2] and Guilford et al. [15]. Two scorers first scored the uses individually, and then compared and discussed scoring differences until a consensus was reached. Every use received between 1 and 6 points. We removed uses that were repeated by the respective dyad/AI (same use for the same object, different or identical formulation) or not comprehensible as invalid.

For **fluency**, each dyad or chatbot received 1 point for each unique use. They did not receive points for answers that explained the same use, in different formulations. However, slight adaptations were accepted (e.g. phone stand and picture stand).

For **flexibility**, the dyads/chatbots received 1 point for each unique category of use that was named (e.g. gardening, playing, cooking). Thus, if multiple uses generated by the same dyad/chatbot fell into the same category of use (as in the phone/picture stand example), they received 1 flexibility point only for the first occurrence of the category, and 0 points for any following uses within the same category, for each AUT object.

For **originality**, we scored 2 originality points if the category of use occurred < 3 times (equivalent to $< 10\%$ of all uses) in the total AUT data (all uses from all dyads and chatbots), 1 point if categories occurred $3 - 5$ times (equivalent to $10 - 20\%$ of all uses), and 0 points if categories occurred > 5 times (in $> 20\%$ of uses).

Finally, for **elaboration**, we scored uses between 0 (no elaboration) and 2 (detailed elaboration) points. As described in Table 2, uses were scored with 0 points when participants gave very short answers without description/method (i.e. How can the object be used/modified to realize this use?), purpose (i.e. Why is the use useful?) or context (i.e. In which context can/would this be used?). Uses were scored with 1 point when they contained some relevant description or context on the use, but it was still not fully comprehensible and uses scored 2 points when they were fully comprehensible and contained relevant context and descriptions.

Table 2: Elaboration Scoring System

Points	Meaning	Definition	Examples
0	No elaboration	One word uses, no explanation, no verb.	Hairbrush.
1	Some elaboration	At least one verb (or relevant context information) and noun, part explanation, not fully comprehensive, missing method or purpose, vague or unspecific.	Use it to brush the cat.
2	Detailed elaboration	More than one word, full and comprehensive explanation, clear description of method, or specific use cases, no unspecified expressions (e.g. something).	Use as comb, if you hold two forks together, you have many more rows and can go through the hair this way.

5 RESULTS

The **fluency** results show very similar *means* in both groups for the overall summed-up fluency score, with a *mean* of 39.500 for the AI group ($md = 39.0$, $sd = 10.50$, $n = 10$), and a *mean* of 39.150 for the human dyads ($md = 37.5$, $sd = 11.43$, $n = 20$). Furthermore, when considering the distributions of both the dyads' and the chatbots' scores, the distributions are relatively uniform and balanced, with approximately one-third of both human dyads and AI chatbots placing in the first, second, and third tertile of the entire group of participants score-wise. Thus, there are good, average, and bad performers in each group, when it comes to creative fluency. However, it must be acknowledged that the fluency score is only limitedly meaningful in this study because we let the AI generate alternate uses in numbers to at least match the human medians. The fact that the AI fluency scores are only marginally higher is due to several occurrences of chatbots repeating uses, or refusing to generate more uses. If we only considered the uses of the AI group generated in response to our initial prompt, which in principle asked the AI - like the human group - to generate "as many alternate uses

as possible", the AI group average would have been much smaller ($mean$ = 26.30, md = 25.5, sd = 10.91).

For the **originality** results, the Mann-Whitney U test reveales the between-group differences to be significant (U = 147.0, $p < 0.05$). Overall, the human dyads came up with more categories of alternate uses than the AI group, which is reflected in higher $means$ of the human group ($mean$ = 23.00, md = 20.5, sd = 9.29), compared with the AI group ($mean$ = 14.50, md = 14.0, sd = 9.90). A visualization of the originality score means is provided in Figure 1.

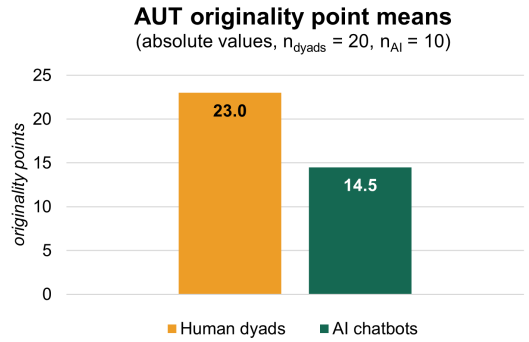

Figure 1: AUT Originality Point Means in Dyads vs. Chatbots

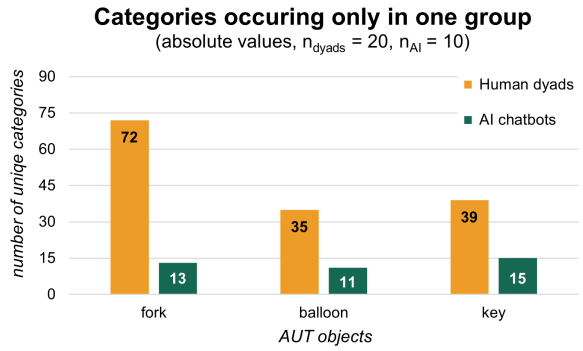

Figure 2: AUT Categories of Use Occurring Only in One Group

As a collective, the human dyads group was therefore distinctly more original than the AI chatbot group. This conclusion is also supported by an analysis of categories that only occur in one experimental group, but not the other. The results of this analysis show that the range of unique categories was much higher in the human group, both in absolute numbers, as depicted in Figure 2, as well as averages ($mean_{\text{dyad}}$ = 7.30 categories, $mean_{\text{AI}}$ = 3.90 categories).

The results in the dimension **flexibility** once again suggest a somewhat stronger performance of human dyads, compared to AI chatbots, though differences were non-significant in the Mann-Whitney-U Test. Human dyads had higher $mean$ points of 34.60 (md = 33.0, sd = 9.32) than the AI group ($mean$ = 31.70, md = 30.0, sd = 8.67), and the human top performers also overscored the AI top performance (max_{dyad} = 52, max_{AI} = 46). Considering all flexibility results from both the human and the AI group, 35% of human dyads scored in the third (top) tertile, and only 25% in the first (bottom) tertile. For the chatbots, 30% scored in the third tertile, but 50% in the first tertile, as visualized in Figure 3.

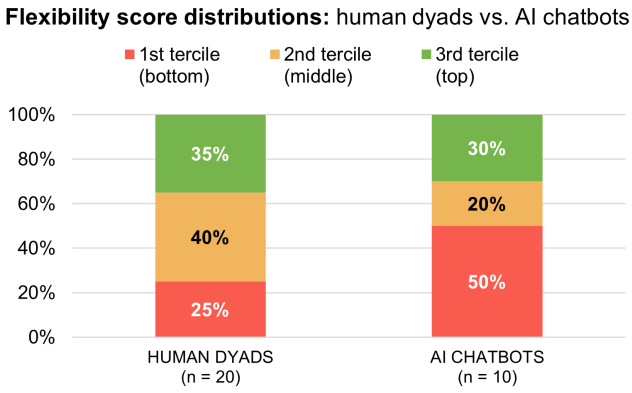

Figure 3: AUT Flexibility Score Distributions into First (Bottom), Second (Middle), and Third (Top) Tertile

These results indicate that, on average, the human dyads came up with a greater range of different alternate uses in the task than the chatbots. One explanation for the lower scores from the chatbots is, that some bots automatically generated multiple uses that belonged to a few specific categories (e.g. cooking, outdoor activities), without such a suggestion from their prompt, which affected their flexibility score.

The results in the AUT score dimension, **elaboration**, show a significantly stronger performance in the AI group (U = 25.0, $p < 0.01$). The chatbots achieved a $mean$ of 76.00 elaboration points, with the human dyads' $mean$ of 48.10 much lower. Elaboration is also the only AUT score dimension in which the highest overall score was achieved by a chatbot (max_{AI} = 107, max_{dyad} = 96).

The summed-up **overall score results** for the AUT, comprised of the score points for fluency, originality, flexibility, and elaboration, are contained in the following Table 3. As detailed in the table, the chatbots' overall score $mean$ is visibly higher than the $mean$ of the human group, which can most likely be attributed to the chatbots' much higher average elaboration scores.

Table 3: AUT Score Sum: Human Dyads vs. AI Chatbots

	mean	*md*	*sd*	*min*	*max*	*n*
Human dyads	144.85	154	41.23	77	240	20
AI chatbots	161.70	162	42.21	89	216	10

The lineup of all participants' scores, as depicted in Figure 4, also shows that five out of the ten AI chatbots are included in the third (top) tertile, though the highest score, in this case, was achieved by a human dyad.

Both the human and the AI group are represented in the first (bottom), second (middle), and third (top) performance-tertile. Individual analyses of chatbot's performances show that Copy AI placed highest out of all AI's in the score dimensions fluency and

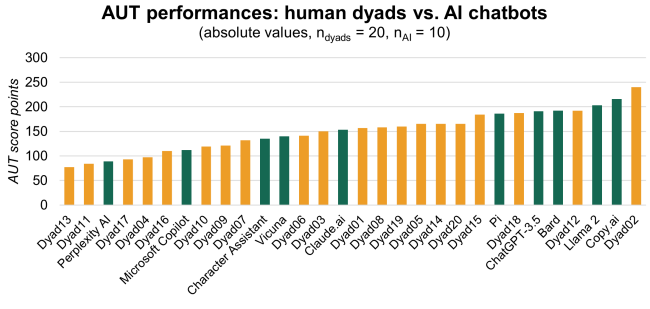

Figure 4: Total Lineup of AUT Performances in the Study

elaboration, Llama2 achieved the highest AI score in flexibility, and ChatGPT-3.5 ranked highest among the chatbots in originality. The top performer from the human dyads was always Dyad02, in all categories but originality. Regarding the other end of the performance spectrum, the bottom performers were consistent throughout all scoring dimensions, for both the AI and the human group.

Considering these results, human dyads generally achieved higher scores *means* in originality and flexibility, while the AI chatbots achieved significantly higher score *means* in elaboration. Emergent observations outside of the AUT scoring system also include that the AI chatbots tended to generate more conventional (e.g. concerning everyday activities), secure (non-dangerous), and realizable (easy to achieve) uses, while humans also came up with unconventional (e.g. concerning limited or very specific contexts of use), riskier (dangerous), and complicated (difficult to achieve) uses.

6 DISCUSSION

In this study, we set out to investigate how different text-generative AI chatbots compare to and differ from human dyads in the Alternate Uses Test (AUT) for measuring creative potential. We assessed 1178 alternate uses, written by 20 human dyads and 10 AI chatbots, in the AUT dimensions of fluency, originality, flexibility, and elaboration. Based on our results, we also derive first insights into understanding if human-AI collaboration could enhance creative performance.

Overall, our results indicate that AI and human creativity indeed differ from each other and excel in different areas. Our **fluency** results show that the selected AI chatbots could generate at least as many alternate uses for different objects as the average human dyad. We chose the median number of uses generated by the dyads as our stopping mechanism for the AI instead of, e.g. time, so we did not challenge the bots to come up with vast numbers of uses. Still, in our study, we experienced repeated answers from three chatbots and had one instance of a bot that was unable to generate more uses at the second prompt. However, given the pace of AI development, we assume that improvements will be achieved regarding AI creative fluency in the foreseeable future.

Results in the **originality** dimension show that humans outrank AI very clearly in originality. Collectively, the human dyads came up with many more unique categories of use than the AI chatbots, indicating that human creativity might have a more individual and less restricted range than AI creativity. These results are consistent with a previous AUT human vs. AI study by Stevenson et al. [32], and also match results from a creativity study by Doshi & Hauser [11], which indicate that humans - as a collective - generate more diverse and novel creative products than AI.

For the **flexibility** dimension, the human dyads portrayed a somewhat enhanced creative performance compared to the AI chatbots. This result aligns with the originality results and indicates that humans tend to think of many different potential uses, while AI sticks to fewer selected categories of use. Again, this observation is consistent with the results from Stevenson et al. [32]. In our small sample, differences in flexibility were non-significant, though more extensive studies might prove otherwise. AI performance might also be improved through training to develop more flexible uses, as the default setting of some AIs seemingly is to write multiple uses within the same category, reducing flexibility.

The results in the **elaboration** dimension show a clear overall superiority of AI, which, on average, elaborated in significantly more words and detail and better justified the usefulness and method of their generated alternate uses. Though part of this distinction might be attributed to the different modes of performing the task (spoken for the human group vs. written for the AI group), the chatbots tended to formulate all uses in complete sentences and explained not only the method but also the purpose of the named alternate use, quite often. On the other hand, participants in the human dyads would often name and not explain a use or vaguely describe methods or purposes. These results are in line with several studies [11, 23, 32], which - taken together - strongly support the assumption that text-generative AI chatbots excel in clear and detailed writing. Following Runco & Jaeger [29], we agree that creative ideation should also include the utility and effectiveness of ideas, which is reflected in the elaboration score.

Overall, our results suggest that human and AI creativity have different strengths and weaknesses. AI seems particularly good at elaborating and coming up with useful and realizable (i.e. easy and safe to carry out) alternate uses. On the other hand, humans tend to focus on generating original and partly unique (e.g. unclear, risky, or very limited purposes of uses, no practicability) uses first and assessing or elaborating on their usefulness only in a second step.[3]

A natural derivation of our results is to investigate the potential of human-AI collaboration for maximizing creative performance. We plan to conduct a comparative experiment on AI-supported vs. unsupported creative collaboration, in which human dyads will once again work on the Alternate Uses Test, with or without support by a state-of-the-art generative AI chatbot. We will examine if and how the chatbot support improves the creative performance in the AUT and explore the effect of AI use on human-to-human collaboration in the study. Based on the results of this preliminary work, we expect general improvements in elaboration for the AI-supported dyads and score elevation for otherwise low-scoring AI-supported dyads. Results from Shaikh & Cruz [31] also suggest that AI might negatively impact the collaboration between human

[3]Please note that the four AUT scoring categories cover different aspects of creativity, but are not independent of each other. For example, flexibility and originality are connected, as they both access the assigned categories of use. But connections may also arise from the data collection process (e.g. time restrictions might lead humans to not elaborate as much) and the differences between human and AI data collection. These circumstances may limit the precision of our results, but cannot be avoided entirely when investigating human vs. AI study subjects.

participants. In future work, it would further be interesting to analyze human vs. AI creative performance in the long term to explore the potential and the potential barriers of AI creativity. This would also provide a more robust measure for human creativity, as human performance can vary day-to-day, based on factors such as stress, nervousness, mental and physical well-being, as well as interest and concentration[23].

It is clear that despite their inferiority to the human group in originality and flexibility, AI chatbots performed very well in this study and often outmatched human dyads. Still, we think it is crucial to be aware that we cannot retrace how AI comes up with answers, and thus, we do not know whether it generates or just copies answers to prompts [23] and that AI answers are always generated from existing data and thus it is a point of ongoing discussion whether or not AI can truly create anything [22].

7 CONCLUSION

This study contributes insights into human vs. AI creative performance and raises questions on the creative potentials of human-AI collaboration. Our results indicate that humans still frequently outperform AI when creating original and diverse ideas and that AI regularly surpasses humans in explaining and justifying generated ideas. As a contribution to the TREW workshop, our results indicate high complementary potential in human-AI creative collaboration, particularly for finding a balance between imagination, precision, and feasibility. Our future studies on human-AI collaboration will investigate this potential while also exploring the social aspects of using AI for collaborative work to establish effective and trustworthy human-AI workflows.

REFERENCES

[1] Spring ACT. 2024. Sophia Chat. https://sophia.chat/

[2] Amin Alhashim, Megan Marshall, Tess Hartog, Rafal Jonczyk, Danielle Dickson, Janet van Hell, Gül Kremer, and Zahed Siddique. 2020. Work in Progress: Assessing Creativity of Alternative Uses Task Responses: A Detailed Procedure. In *2020 ASEE Virtual Annual Conference Content Access*. ASEE, Virtual On line. https://doi.org/10.18260/1-2--35612

[3] Mayssa Ahmad Ali Elfa and Mina Eshaq Tawfilis Dawood. 2023. Using Artificial Intelligence for enhancing Human Creativity. *Journal of Art, Design and Music* 2, 2 (2023), 3.

[4] Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. Which Humans? https://doi.org/10.31234/osf.io/5b26t

[5] Christian Bergmann and Ferdinand Eder. 2005. *AIST-R. Allgemeiner-Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R). Revision.* Beltz Test Gesellschaft, Göttingen, Germany.

[6] Francisco Castro, Jian Gao, and Sébastien Martin. 2023. Human-AI Interactions and Societal Pitfalls. https://doi.org/10.48550/ARXIV.2309.10448 Publisher: arXiv Version Number: 2.

[7] Nicholas Davis, Chih-Pin Hsiao, Yanna Popova, and Brian Magerko. 2015. An Enactive Model of Creativity for Computational Collaboration and Co-creation. In *Creativity in the Digital Age*, Nelson Zagalo and Pedro Branco (Eds.). Springer London, London, UK, 109–133. https://doi.org/10.1007/978-1-4471-6681-8_7 Series Title: Springer Series on Cultural Computing.

[8] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, Sanat Moningi, and Brian Magerko. 2015. Drawing apprentice: An enactive co-creative agent for artistic collaboration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (Glasgow, United Kingdom). Association for Computing Machinery, New York, NY, USA, 185–186. https://doi.org/10.1145/2757226.2764555

[9] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, and Brian Magerko. 2016. Empirically Studying Participatory Sense-Making in Abstract Drawing with a Co-Creative Cognitive Agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma California USA, 2016-03-07). ACM, New York, NY, USA, 196–207. https://doi.org/10.1145/2856767.2856795

[10] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal* 013, 24 (2023). https://doi.org/10.2139/ssrn.4573321

[11] Anil R. Doshi and Oliver P. Hauser. 2023. Generative artificial intelligence enhances individual creativity but reduces the collective diversity of novel content. *CoRR* abs/2312.00506 (2023). https://doi.org/10.48550/ARXIV.2312.00506 arXiv:2312.00506

[12] Drift. 2024. Drift | Everything Starts With a Conversation. https://www.drift.com/

[13] GitHub. 2024. GitHub Copilot. https://github.com/features/copilot

[14] Joy Paul Guilford. 1967. *The nature of human intelligence*. Vol. 1. McGraw-Hill, USA.

[15] Joy P. Guilford, Paul R. Christensen, Philip R. Merrifield, and Robert C. Wilson. 1960, 1978. *Alternate Uses Manual. Menlo Park.* CA: Mind Garden, Menlo Park, CA, USA.

[16] Vivian Emily Gunser, Steffen Gottschling, Birgit Brucker, Sandra Richter, Dîlan Canan Çakir, and Peter Gerjets. 2022. The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors?. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)* (Dublin, Ireland). Association for Computational Linguistics, Dublin, Ireland, 60–61. https://doi.org/10.18653/v1/2022.in2writing-1.8

[17] Jennifer Haase, Djordje Djurica, and Jan Mendling. 2023. The Art of Inspiring Creativity: Exploring the Unique Impact of AI-generated Images. In *AMCIS 2023 Proceedings* (United States), Paul A. Pavlou, Vishal Midha, Animesh Animesh, Traci A. Carte, Alexandre R. Graeml, and Alanah Mitchell (Eds.). Association for Information Systems, Panama City, Panama, 1–10.

[18] Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. 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 (2023), 107502. https://doi.org/10.1016/j.chb.2022.107502

[19] Jess Hohenstein and Malte Jung. 2020-05. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020-05), 106190. https://doi.org/10.1016/j.chb.2019.106190

[20] John L Holland. 1997. *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources, USA.

[21] Lu Hong and Scott E. Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389. https://doi.org/10.1073/pnas.0403723101

[22] Keith Kirkpatrick. 2023. Can AI Demonstrate Creativity? *Commun. ACM* 66, 2 (2023), 21–23. https://doi.org/10.1145/3575665

[23] Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports* 13, 1 (2023), 13601. https://doi.org/10.1038/s41598-023-40858-3

[24] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, 2020-04-21). ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376739

[25] Mary Lou Maher. 2012. Computational and Collective Creativity: Who's Being Creative?. In *Proceedings of the Third International Conference on Computational Creativity*. International Conference on Computational Creativity 2012, Dublin, Ireland, 67–71. https://api.semanticscholar.org/CorpusID:17655886

[26] OpenAI. 2024. ChatGPT. Get instant answers, find creative inspiration, learn something new. https://openai.com/chatgpt

[27] Jeba Rezwana and Mary Lou Maher. 2023. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 67 (sep 2023), 28 pages. https://doi.org/10.1145/3519026

[28] Mark A Runco and Selcuk Acar. 2012. Divergent thinking as an indicator of creative potential. *Creativity research journal* 24, 1 (2012), 66–75.

[29] Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal* 24, 1 (2012), 92–96. https://doi.org/10.1080/10400419.2012.650092 arXiv:https://doi.org/10.1080/10400419.2012.650092

[30] R. Keith Sawyer and Stacy DeZutter. 2009. Distributed creativity: How collective creations emerge from collaboration. *Psychology of Aesthetics, Creativity, and the Arts* 3, 2 (05 2009), 81–92. https://doi.org/10.1037/a0013282

[31] Sonia Jawaid Shaikh and Ignacio F Cruz. 2023. AI in human teams: effects on technology use, members' interactions, and creative performance under time scarcity. *AI & SOCIETY* 38, 4 (2023), 1587–1600.

[32] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting GPT-3's Creativity to the (Alternative Uses) Test. arXiv:2206.08932 [cs.AI]

[33] Henrik Skaug Sætra. 2023. Generative AI: Here to stay, but for good? *Technology in Society* 75 (2023), 102372. https://doi.org/10.1016/j.techsoc.2023.102372

[34] Evan Weingarten, Michael W. Meyer, Amit Ashkenazi, and On Amir. 2020. Human Experts Outperform Technology in Creative Markets. *She Ji: The Journal of Design, Economics, and Innovation* 6, 3 (2020), 301–330. https://doi.org/10.1016/j.sheji.2020.07.004

[35] Lauren Winston and Brian Magerko. 2017. Turn-taking with improvisational co-creative agents. In *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Little Cottonwood Canyon, Utah, USA, 2017) *(AIIDE'17)*. AAAI Press, New York, NY, 129–135.

[36] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

Received 22 February 2024; revised 17 April 2024; accepted 17 April 2024