scientific reports

OPEN

Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence

Jack McGuire[1,3]✉, David De Cremer[1,3] & Tim Van de Cruys[2]

The emergence of generative AI technologies has led to an increasing number of people collaborating with AI to produce creative works. Across two experimental studies, in which we carefully designed and programmed state-of-the-art human–AI interfaces, we examine how the design of generative AI systems influences human creativity (poetry writing). First, we find that people were most creative when writing a poem on their own, compared to first receiving a poem generated by an AI system and using sophisticated tools to edit it (Study 1). Following this, we demonstrate that this creativity deficit dissipates when people co-create with—not edit—AI and establish creative self-efficacy as an important mechanism in this process (Study 2). Thus, our findings indicate that people must occupy the role of a co-creator, not an editor, to reap the benefits of generative AI in the production of creative works.

Keywords Creativity, Co-creation, Artificial intelligence, Self-efficacy

Current developments in large language models (LLMs) have led to a surge of interest in the application of generative AI technologies in creative domains[1–6]. Recent research has largely skewed in favour of touting the significant opportunities associated with human–AI collaboration, where generative AI has been found to exhibit human-like levels of creativity[7,8]. Given this track record for exhibiting creativity in a selection of narrow domains, scholars have been quick to reason that generative AI can also augment and enhance human creativity[9,10]. However, research on human–AI creative collaboration is surprisingly limited. In this research, we therefore set out to improve our understanding of the factors that drive effective creative collaboration with AI.

First, we begin with the observation that generative AI has meaningful limitations in its production of creative work, which suggests that publicity surrounding its creative prowess may not yet map onto reality. For example, seasoned writers have been found to outperform (by a large margin) current LLMs (e.g., Claude, GPT-4) across a diverse range of creativity measures[11]. When judging a short story produced by GPT-4, an expert writer observed that "the events of this piece feel arbitrary, almost random" and "while that does grant it an unpredictability and a vague form of originality, it feels thoughtless"[11, p. 20]. In addition, AI generated creative writing has been found to be repetitive, following a homogenous form and writing style[12,13], further contributing to its lack of originality and novelty[14]. This particular point becomes evident when comparatively evaluating large bodies of AI generated work, where similarities become obvious.

In addition, there are important omissions and limitations with the present literature on human-AI creative collaboration. A first limitation is a methodological one. Specifically, a great deal of empirical research that has sought to test the creative capabilities of LLMs have not used "expert" evaluators (e.g., domain experts evaluate creative outcomes). Instead, this body of work has made use of "lay" evaluations of creativity (laypeople evaluate creative outcomes) to substantiate claims of creative generative AI[7,8,15,16]. This is important because obtaining evaluations from experts is widely considered the "gold standard" when evaluating creative outcomes[17,18]. Furthermore, although several studies have compared AI generated creativity against a human benchmark[8,16], a surprisingly limited amount of work has examined the creative output of human-AI dyads and the collaborative processes that undergird this. Hence, there is still much to be known about how AI systems should be designed to best support human creativity.

[1]Department of Management & Organizational Development, D'Amore-McKim School of Business, Northeastern University, Hayden Hall, 101, 370 Huntington Ave, Boston, MA 02115, USA. [2]Linguistics Research Unit, Faculty of Arts, KU Leuven, Blijde-Inkomststraat 21, Box 3308, 3000 Leuven, Belgium. [3]Department of Management and Organisation, National University of Singapore, Singapore, Singapore. ✉email: jo.mcguire@northeastern.edu

Finally, of the limited body of work that has researched the creativity of human–AI dyads (e.g.[15]), most lack the internal validity of controlled and robustly designed experiments. An illustrative example of a typical approach here is to provide participants with instructions on a set of prompts to provide GPT-4 and request that the creative output derivative of this process is inputted back into the experimental interface (e.g., Qualtrics). Such an approach, however, poses a clear threat to internal validity because the experimenter has little to no control over participant adherence to instructions[19]. Moreover, the evolving nature of publicly available LLMs presents further challenges because identical prompts at different points in time can produce very different results[20] and changes to model specifications can lead to fluctuations in performance, even across short periods of time[21]. Thus, as many scholars seek to utilize widely accessible and publicly available LLMs to conduct research on the creativity of human–AI dyads, they also invite important threats to the validity of their findings.

We assume that the way people view their role in creative collaborations has an influence on the outcome of their endeavours[22]. This assumption therefore implies that how generative AI is utilized, as such affecting the role humans occupy in the creative process, will be highly consequential for creative outcomes. In the present research, we focus on the role of people serving either as a co-creator or an editor. The occupation of these two roles is expected to have a significant influence on people's experience of creative self-efficacy in human-AI collaboration. When people are placed in the role of an editor, as is often the case by default in human-AI collaboration[2,23], AI determines the initial direction of the creative process because it provides a default creative product that the editor must evaluate. This will likely pose a threat to the editor's experience of self-efficacy because AI occupies the proactive role of generating a creative product and the editor occupies the reactive role of aiding with improving its work[24,25]. Moreover, placing people in the role of an editor can invoke an anchoring effect, where 'editors' are highly influenced by the default they are presented with and make insufficient edits as a consequence[15]. Such constraints, whether consciously or unconsciously processed, may also lead to insufficient editing because they undermine intrinsic motivation, an important antecedent to creative processes[26].

Alternatively, when people are placed in the role of a co-creator, this is similar to a Cyborg archetype of human–AI collaboration, where people work in tandem with AI rather than divide the labour between the two[27]. When occupying this co-creator role, peoples' belief in their ability to produce creative outcomes is nurtured because they set the creative direction, rather than simply react to AI-generated output[28]. In addition, opportunities for spontaneous improvisation are preserved when the creative product is developed iteratively because possibilities remain open[29]. Such facets of co-creation will likely nurture intrinsic motivation[26], which is constructive for the development of self-efficacy[30]. We therefore argue that human–AI creativity will be best fostered when humans are placed in the role of a co-creator (vs. an editor) and this will be explained by the promotion of creative self-efficacy. Finally, as poetry embodies the most prototypical expression of creativity—spanning centuries of human history[31,32]—and enables us to test the effect of being a co-creator vs. editor on creative outcomes, this research focuses on poetry writing as the medium of creative expression.

We conducted two rigorously designed experiments in which we carefully designed and programmed state-of-the-art human–AI interfaces. Specifically, we find that people are less creative (as judged by professional poets) in a poetry writing task when they collaborate with a generative AI system and are placed in the role of an editor, relative to writing a poem on their own (Study 1). Interestingly, however, we further find that this creativity deficit disappears when people are instead placed in the role of a co-creator and this effect is explained by greater reported levels of creative self-efficacy (Study 2). Both experiments were approved for use with human subjects by an institutional review board. All conditions, measures and exclusions are reported; data are available at https://osf.io/3jqga/.

Results

Study 1

Participants were 101 individuals recruited via Prolific Academic in exchange for GBP2.50 (approximately USD3.00). Six participants failed to correctly answer an instrumental attention check and were removed from subsequent analyses (see below). Of the remaining 96 participants, on average, they were 27.23 years old (SD = 8.97) and 76% were female. Participants were randomly allocated to one of two experimental condition groups. In the first condition (human condition; n = 48), participants completed a poetry task unassisted. In the second experimental condition (human–AI condition; n = 48), participants completed a poetry task in collaboration with an artificially intelligent poetry generation system. A third experimental condition (AI condition) was added where the poetry task was completed entirely by the artificially intelligent poetry generation system (n = 50).

Participants were invited to take part in a study about creativity and were informed they would respond to some questions regarding their experiences of completing a creative task. The study was accessible via a URL link that was shared on our Prolific study advertisement. For participants in both conditions, the study begins with a welcome message, followed by a poetry task. Participants allocated to the human condition were provided with the following instructions:

> The first half of our study will involve writing a poem.
> For this study, you will be asked to produce a poem that is 8 lines in length, consisting of two stanzas that each have 4 lines (2 × 4 lines). A stanza in poetry is like a paragraph in prose writing.
> The rest is up to you. It's entirely up to you whether you want to make it rhyme or not.
> There will be no time limit, please do take whatever time you need to write your poem.

Participants were then asked whether they understood the instructions for the task (yes/no) and given no time limit to write their poem. All participants in this condition selected 'yes'. The instructions for participants in the human–AI condition were more elaborate. In addition to being told that they will be asked to write a poem

8 lines in length (two stanzas, four lines each), they were also told that they will first be provided with a poem produced by an algorithm—a set of instructions that are carried out in a series of computations/calculations, typically performed by a machine—and that they can edit the poem as freely as they wish before submitting it and finishing the task. More specifically, we instructed participants:

> Once you are presented with the novel poem produced by the algorithm, you are free to edit and change the poem in whatever way you want. This part is entirely up to you. You can leave the poem as it is, make only a few minor adjustments, or delete the poem and start from scratch.

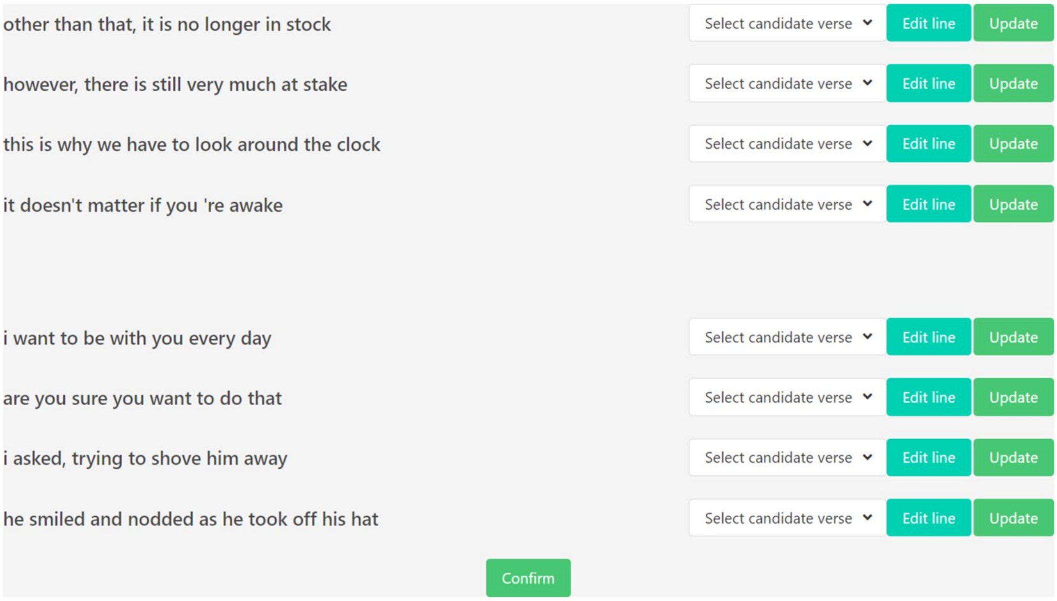
The algorithm used to generate novel poetry is a state of the art, scientifically validated, neural poetry generation system that has been shown to generate output comparable to human-written poetry[33]. The poetry system was trained exclusively on a corpus of standard, nonpoetic text—derived from the CommonCrawl corpus (https://commoncrawl.org/). After a series of filtering rules were applied to this vast and generic corpus of text (e.g. retain only sentences written in English that are 20 words or less), the resulting corpus from which the poetry generation system was trained on contains 500 million words in total, which is constructed from a vocabulary of 15 thousand unique words. The poetry generation system uses a recurrent neural encoder-decoder architecture in order to generate individual lines of poetry. In total, 2048 potential candidate lines are generated by the poetry system for consideration to be included in each line. The candidates generated are subject to rhyme and topical constraints. The rhyme constraint imposed by the system follows an ABAB rhyme structure (e.g. lines 1 and 3 rhyme, and lines 2 and 4 rhyme). Therefore, potential lines generated by the system for line 3 must rhyme with the line chosen for line 1. The topical constraint ensures that the poem generated by the system is topically coherent. For this, the system makes use of a latent topic model based on non-negative matrix factorization[34]. The topic model provides interpretable, topically coherent semantic dimensions, which are exemplified by the three most salient words. In short, these three topical words represent the 'theme' of the poem. Examples of combinations of three topical words are "sorrow, longing, admiration" and "goddess, ritual, shrine". For this study, the system generates a poem based on one set of three topical words and this set is randomly selected from a list of 100 sets of three topical words. As the generation of potential lines is a sampling process, the sample of lines generated by the system are then subject to a global optimization framework to identify the line with the best match according to the constraints mentioned above. To rank order the sample of potential lines, each line is evaluated according to its compliance with the rhyme structure, its compliance with the topical constraint, the number of syllables it contains, and the log-probability score of two further mathematical criteria[35]. The line with the highest aggregated score across these dimensions is then selected for inclusion in the poem. This process is followed for all eight lines of each poem generated. Evaluations of the quality of poems produced by this poetry generation system revealed that raters considered the poems to be comparable to works by well-established English poets (W.H. Auden, E.E. Cummings, Philip Larkin, Sarojini Naidu, and Sylvia Plath), and superior to previous poetry systems (Hafez and Deep-speare) across dimensions of fluency, coherence, meaningfulness, and poeticism. This indicates the poetry generation system we have adopted for this research is state-of-the-art.

The interface for the assistive poetry generation system we incorporated in our study was modified and designed to include several features that can support the creative process of poetry writing and provide participants with a high degree of freedom. This is because the original poetry generation system was designed to automate poetry generation and not augment human poetry writing. Therefore, we developed an interface that builds upon this existing poetry generation system[33] by implementing design changes with the goal to assist and facilitate human creativity (full details regarding the required back- and front-end development are provided in the Methods section). We described to participants in detail what each of these features are prior to the poetry task beginning. The first feature is the ability for participants to directly edit any line in the poem. The second feature is a dropdown function that is available for each line and contains alternative lines that participants can choose from if they prefer an alternative line to an existing line. A third feature provided participants with an updated list of alternative lines, if they directly edit and update one of the lines that they wish to browse alternative lines for. In this way, the dropdown list of alternative lines updates in real-time based on the newly edited line. Finally, as the poetry system generates poems that follow an ABAB rhyme scheme (e.g. lines 1 and 3 rhyme, and lines 2 and 4 rhyme), if a participant edits and updates lines 1 or 2, the updated alternative lines for lines 3 or 4 will rhyme with the newly edited line (see Fig. 1 for a visual illustration of the poetry interface). The goal of designing the interactive poetry system in this way was to provide participants with assistive, artificially intelligent input for the generation of their own poems. As in the human condition, participants were reminded that no time limit would be imposed and that they could spend however long they wished to write and submit their poem. All participants in this condition responded 'yes' when asked whether they understand the instructions of this poetry task (yes/no).

Once participants submitted their finished poems, they subsequently provided their self-reported evaluation of the poem's creativity.

To measure participant's self-reported evaluations of the creativity of their poem, we instructed: "rate the extent to which you think the poem you submitted is creative" (1 = lowest creativity score; 40 = highest creativity score)".

To evaluate the creativity of participant's poems, we utilized the Consensual Assessment Technique (CAT), one of the most highly regarded assessment tools in creativity research[17]. In the CAT, evaluations of creativity are obtained from experts who are asked to evaluate creative products using their own expert sense of what is creative. We obtained evaluations from 10 poets in exchange for USD40 each. All poets we obtained evaluations from are professional writers and have had their poems published in either their own poetry books or as part of a poetry anthology that contains works from different poets. In addition, 6 of the poets possess a graduate degree (MA, MFA, PhD) in creative writing, English literature, or a related field. Poets were asked to evaluate



other than that, it is no longer in stock
however, there is still very much at stake
this is why we have to look around the clock
it doesn't matter if you 're awake

i want to be with you every day
are you sure you want to do that
i asked, trying to shove him away
he smiled and nodded as he took off his hat

Figure 1. Visual depiction of the poetry interface in the human-AI condition (Study 1).

the creativity of a collection of poems. This collection contained all the poems produced by participants in the human condition (n = 48), the human-AI condition (n = 48) and the poems generated in the AI condition (n = 50). Therefore, in sum, each poet evaluated the same set of 146 poems. Importantly, poets were not provided any information about the authors behind the poems or the circumstances under which they were written. We included poems generated autonomously by the poetry generation system to obtain a benchmark value for comparison with creativity ratings obtained in our two experimental conditions. In line with creativity research that utilizes the CAT[35–37], poets were given the poems in different, randomly generated orders and asked to rate the creativity of each poem on a 1 (lowest creativity score) to 40 (highest creativity score) scale. They were also informed that their ratings of poems should be relative to one another, that they should use the full range for their scores (1–40), and that they would not have to justify or defend any of their ratings.

Independent samples t-tests reveal that participants in the human condition were significantly more likely to self-report their poems as more creative (M = 23.88, SD = 8.40), when compared to participants in the human–AI condition (M = 19.13, SD = 10.40, p = 0.016, Cohen's D = 0.47). Next, we analysed how creative the poems were judged to be by expert evaluators (poets). A one-way ANOVA revealed that the creativity of the poems differed significantly across the three experimental groups, F(2, 145) = 63.48, p < 0.001, partial η2 = 0.47 (see Table 1). We then performed post hoc pairwise comparisons to compare expert evaluations of creativity across conditions. Poems in the human condition (M = 18.23, SD = 4.94) were regarded as more creative than poems in the human-AI condition (M = 12.55, SD = 4.32; p < 0.001) the AI condition (M = 9.45, SD = 2.00; p < 0.001), while poems in the human-AI condition were regarded as more creative than poems in the AI condition (p < 0.001).

Study 2

Study 2 builds on the findings of Study 1 by testing whether a redesigned poetry generation system—one that fosters creative self-efficacy—would improve expert evaluations of creativity. We predicted that participants would exhibit greater creativity when placed in the role of a co-creator (co-creator human–AI condition), relative to being placed in the role of an editor (human–AI condition), and this effect would be explained by greater perceptions of creative self-efficacy.

Participants were 152 individuals recruited via Prolific Academic in exchange for GBP2.50 (approximately USD3.00). No participants failed to correctly answer an instrumental attention check. On average, participants were 35.11 years old (SD = 12.06) and 34.9% were female. Participants were randomly allocated to one of three experimental condition groups. In the first condition (human condition; n = 51), participants completed a poetry

	AI condition	Human-AI condition	Human condition
Self-reported creativity	–	19.13$_a$ (10.40)	23.88$_b$ (8.40)
Expert evaluations of creativity	9.45$_a$ (2.00)	12.55$_b$ (4.32)	18.23$_c$ (4.94)

Table 1. Variable means and standard deviations by condition (Study 1). Values in brackets refer to standard deviations. Within a row, values with different subscripts are significantly different (p < 0.05) based on t-tests and LSD tests. *p < 0.05, **p < 0.01, ***p < 0.001.

task unassisted and in the second experimental condition (human-AI condition; n = 50), participants completed a poetry task in collaboration with the same artificially intelligent poetry generation system used in Study 1. A third experimental condition (co-creator human-AI condition) was included where participants completed a poetry task in collaboration with a redesigned version of the poetry generation system (n = 51).

The study followed the same procedure as Study 1. However, we also examined the effects of participants collaborating with a new poetry generation system that was designed to support creative self-efficacy (co-creator human–AI condition. In the co-creator human–AI condition, participants collaborated with the same poetry generation system used in Study 1[33], however, the interface was redesigned so that participants and the poetry system would take turns writing one line of poetry each in a step-wise fashion. In other words, this collaborative and iterative process was akin to a conversation, where the participant begins the conversation by writing the first line of poetry, which is then followed by the AI system generating the next line, and so forth (see Fig. 2 for a visual depiction). In addition, to ensure that the system generates subsequent lines of poetry that align with the direction the participant wishes to take the poem, we had participants select a set of three topical words (e.g., "sorrow, longing, admiration") from a list of 100 sets of three topical words. Thus, participants were also empowered to determine the "theme" of the poem by choosing their topical words before initiating the writing process. This differs from Study 1, where the poetry generation system produced a poem that was based on a randomly selected set of three topical words. Similar to Study 1, the poetry system in this condition followed an ABAB rhyme scheme, where the 4th (or 8th) line generated would rhyme with the 2nd (or 6th) line. It was also made clear to participants in this condition that they were free to edit all and any aspects of the poem at any stage. Thus, whereas the original poetry generation interface (human-AI condition) provided participants with a finished poem and a selection of advanced features to edit this poem, the redesigned interface (co-creator human–AI condition) empowered users to initiate the creative process by selecting their set of topical words and writing the first line of poetry. In turn, the system responded to each user-generated line by returning a subsequent line. Further building on Study 1, we also measured participant's perceptions of creative self-efficacy as a potential mediating mechanism, in addition to their self-reported evaluations of the poem's creativity.

To measure creative self-efficacy, we adopted a three-item ($\alpha = 0.92$) creative self-efficacy scale developed by Tierney and Farmer[28]. The scale captured to what extent participants harboured the belief that they possessed the capacity to perform creative work effectively (e.g., "I was good at coming up with ideas for the poem"). The scale ranged from 1 (strongly disagree) to 7 (strongly agree).

To measure participant's self-reported evaluations of the creativity of their poem, we instructed: "rate the extent to which you think the poem you submitted is creative" (1 = lowest creativity score; 40 = highest creativity score)".

To evaluate the creativity of participant's poems, we utilized the Consensual Assessment Technique (CAT), one of the most highly regarded assessment tools in creativity research[17]. As in Study 1, we obtained ratings from 10 (different) poets in exchange for USD40 each.

Creative self-efficacy

A one-way ANOVA revealed that participants' creative self-efficacy differed significantly across the three experimental conditions F(2, 149) = 7.01, p = 0.001, partial η2 = 0.09 (see Fig. 3). We then performed post hoc pairwise comparisons to compare creative self-efficacy across conditions. Participants in the human condition (M = 4.71, SD = 1.45) reported greater creative self-efficacy than participants in the human–AI condition (M = 3.74, SD = 1.43; p = 0.001). Creative self-efficacy reported by participants in the co-creator human–AI condition (M = 4.62, SD = 1.43) did not differ significantly from participants in the human condition (p = 0.755) but was significantly greater than creative self-efficacy reported in the human–AI condition (p = 0.003).

Self-reported creativity

Next, a one-way ANOVA revealed that self-reported creativity differed significantly across the three experimental conditions F(2, 149) = 6.08, p = 0.003, partial η2 = 0.08. We then performed post hoc pairwise comparisons to compare self-reported creativity across conditions. Participants in the human condition reported greater self-reported creativity than participants in the human-AI condition (M = 23.29, SD = 9.24 vs. M = 16.96, SD = 9.56; p = 0.001). Self-reported creativity reported by participants in the co-creator human-AI condition (M = 21.16,

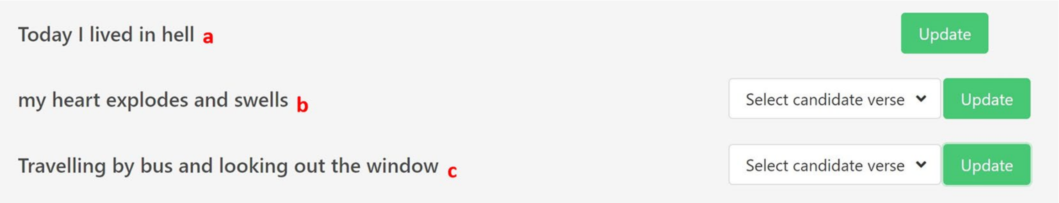

Today I lived in hell
my heart explodes and swells
Travelling by bus and looking out the window

Figure 2. Visual depiction of the poetry interface in the co-creator human-AI condition (Study 2). The participant begins by writing the first line of the poem (**a**). After selecting the 'update' button on the right-hand side, the poetry generation system returns the second line of the poem (**b**). The option to select alternative lines is reflected in the feature 'select candidate verse' and participants can directly edit the line by clicking on the line. Once satisfied, the participant can write the third line of the poem in the entry below (**c**). This iterative process continues until the participant co-creates an 8-line poem consisting of 2 stanzas (2 × 4 lines).

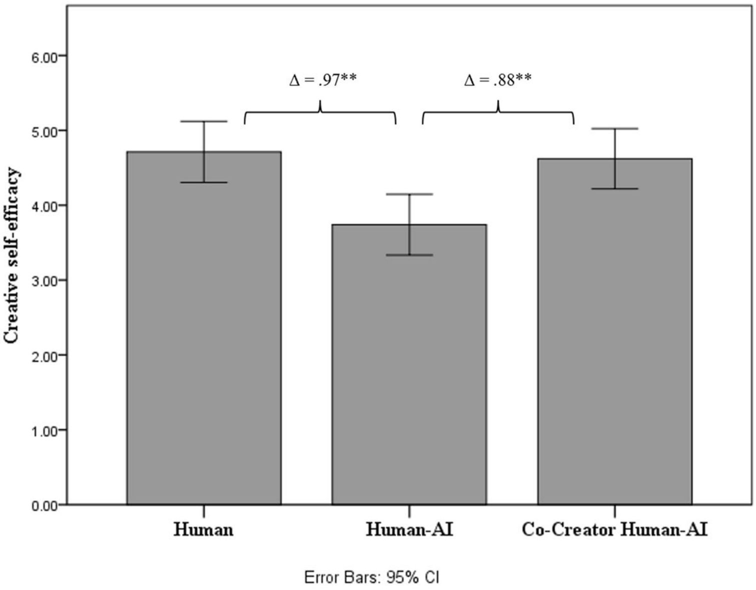


Figure 3. Creative self-efficacy reported across experimental conditions (Study 2).

SD = 9.03; p = 0.246) was also significantly higher than participants in the human-AI condition (p = 0.025), however, it did not differ significantly from participants in the human condition (p = 0.247).

Expert evaluations of creativity

A further one-way ANOVA revealed that expert evaluations of creativity differed significantly across the three experimental conditions F(2, 149) = 5.77, p = 0.004, partial η2 = 0.07 (see Fig. 4). We then performed post hoc pairwise comparisons to compare expert evaluations of creativity across conditions. Participants in the human condition received greater expert evaluations of creativity than participants in the human–AI condition (M = 15.74, SD = 9.24 vs. M = 12.53, SD = 4.45; p = 0.001). Expert evaluations of creativity received by participants in the co-creator human–AI condition (M = 14.70, SD = 5.06; p = 0.280) were significantly greater than participants in the human–AI condition (p = 0.026), however, these evaluations did not differ significantly from participants in the human condition (p = 0.278). See Table 2 for cell means and standard deviations.

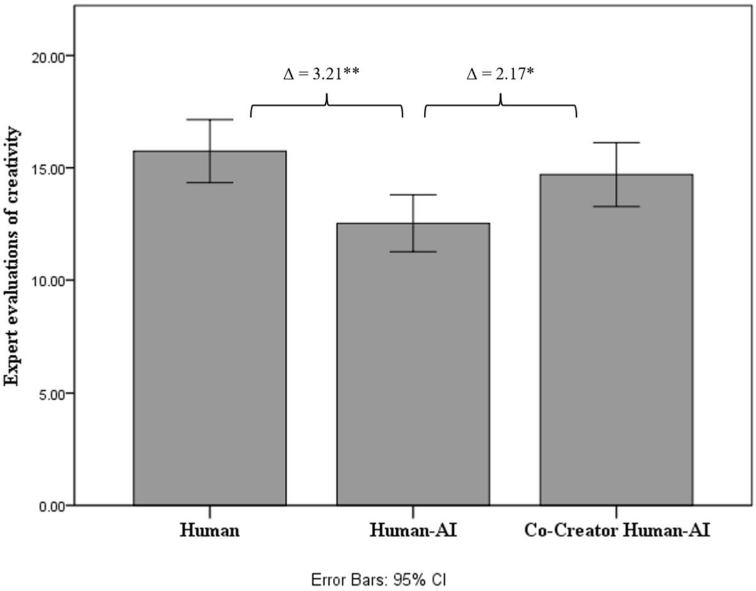


Figure 4. Expert evaluations of creativity across experimental conditions (Study 2).

	Human-AI condition	Co-creator human-AI condition	Human condition
Creative self-efficacy	3.74_a (1.43)	4.62_b (1.43)	4.71_b (1.45)
Self-reported creativity	16.96_a (9.56)	21.16_b (9.03)	23.29_b (9.24)
Expert evaluations of creativity	12.53_a (4.45)	14.70_b (5.06)	15.74_b (4.98)

Table 2. Variable means and standard deviations by condition (Study 2). Values in brackets refer to standard deviations. Within a row, values with different subscripts are significantly different ($p < 0.05$) based on t-tests and LSD tests. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Mediation analyses

Next, we conducted mediation analyses by looking at whether creative self-efficacy mediated the link between being a co-creator (co-creator human–AI condition) versus being an editor (human–AI condition) in creative collaboration with AI and expert evaluations of creativity. To do this, we used Model 4 of the Process macro[38] with 95% bias-corrected confidence intervals and bootstrapped samples set to 10,000. In this path model, creative self-efficacy was entered as a mediator and expert evaluations of creativity was entered as the dependent variable (Y). For the independent variable, we created a dummy-coded variable using indicator coding: the human–AI condition was given the value 0, and the co-creator human-AI condition was given the value 1. A significant indirect effect was observed from our dummy variable to expert evaluations of creativity via creative self-efficacy: indirect effect, B = 0.78, SE = 0.39, 95% CI [0.16, 1.68]. In addition, we also performed the same mediation analyses with a dummy-coded independent variable where co-creator human–AI condition was given the value 0, and the human condition was given the value 1. For this model, the indirect effect was not significant: indirect effect, B = 0.09, SE = 0.29, 95% CI [− 0.50, 0.78] (see Fig. 5).

Discussion

Taken together, our findings demonstrate the importance of how AI systems are designed and the role humans are intended to serve in the production of creative goods. We find that creative collaboration with AI is most effective when AI systems are designed to nurture end-users' beliefs in their abilities to produce creative outcomes. More specifically, we find that this is accomplished through placing people in the role of a co-creator, rather than an editor. Our finding aligns fully with the overarching goal of effective human–computer interaction (HCI): to create intuitive and empowering interfaces that seamlessly integrate with human capabilities[39,40].

Our results run counter to recent suggestions and beliefs espoused by technology leaders and scholars that generative AI can promote creativity by having AI generate outputs and people provide the final judgment on the quality of these outputs[41–43]. Or, put differently, when people are placed in the role of an "editor". This is because preservation of autonomy is crucial for promoting peoples' beliefs about their ability to produce creative products[44,45]. When occupying the role of an editor, the presentation of a default body of text for editing restricts autonomy because people are unconsciously influenced by default effects[15,46]. According to role theory[47,48], occupying the role of an editor will also implicitly set expectations about the behaviours the role entails. Thus, the

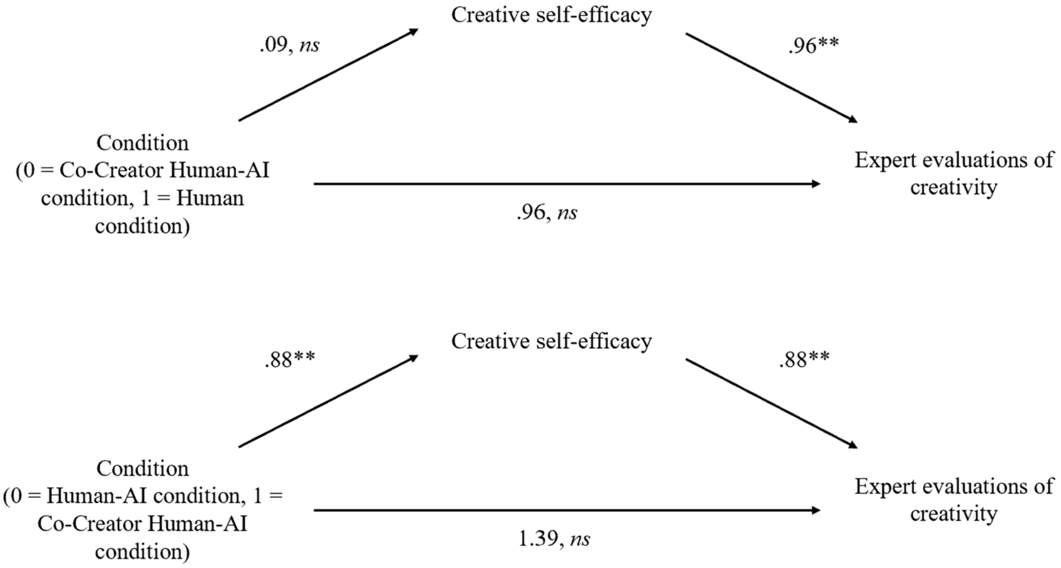

Creative self-efficacy
.09, ns
.96**
Condition (0 = Co-Creator Human-AI condition, 1 = Human condition)
Expert evaluations of creativity
.96, ns
Creative self-efficacy
.88**
.88**
Condition (0 = Human-AI condition, 1 = Co-Creator Human-AI condition)
Expert evaluations of creativity
1.39, ns

Figure 5. Mediation analyses of the effect of our experimental conditions on expert evaluations of creativity via creative self-efficacy (Study 2). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

implicit assumption of a rigidly defined role will undermine self-efficacy as people will interpret their role as not requiring creative and explorative behaviours. In turn, this inhibits cognitive flexibility and divergent thinking[49]. Alternatively, assuming the role of a co-creator will give rise to expectations that one should fully express their ideas, experiment with different possibilities, and challenge themselves[50,51]. Hence, the roles given to people in creative collaboration with AI has a large influence on their self-efficacy and resulting creative contributions.

From a cognitive sciences vantage point, our research also highlights an interesting array of future research directions that relate to the cognitive processes underpinning creative collaboration with AI. For instance, when people occupy the role of a co-creator, this might activate broader associative networks and broaden cognitive focus[52–54]. Alternatively, occupying the role of an editor could activate analytical thinking and elicit a narrower cognitive focus[55,56]. The production of creative ideas has been found to be poorer in virtual teams, when compared to face-to-face teams, because it focuses people's attention to a screen and narrows their cognitive focus[57]. Similarly, when people edit AI-generated content, their cognitive focus likely narrows in on the output generated. Future research could delve into the specific cognitive mechanisms activated during these roles, such as differences in attention ("does editing influence eye gaze?"), decision making style ("does co-creating promote spontaneous improvisation and risk taking?") or affect ("does editing heighten cognitive load and induce stress?"). Finally, longitudinal studies could explore the long-term cognitive effects of consistent collaboration with AI on creative tasks, potentially identifying shifts in cognitive processes over time. For example, if persistent collaboration with AI habituates analytical thinking and a narrower cognitive focus, it would be worthwhile to examine whether this affects creative processes even in the absence of AI (deskilling effect).

Our findings have important implications for the way practitioners understand human-AI interaction in the context of creative collaboration. First, in creative industries such as advertising, design, and content creation[58], implementing AI systems that emphasize co-creation can ensure that organizations tap into the strengths of both human capital and generative AI, though caution should be exercised to prevent deceptive practices[59]. This approach can empower employees and foster a more engaging and satisfying work environment[60]. Similarly, in education, these insights have implications for how AI can be leveraged to maintain and cultivate students' creative self-efficacy[61]. For instance, in creative writing classes, AI tools designed to place students in the role of co-creator could help them develop their writing skills by providing iterative feedback and suggestions, thereby enhancing their self-efficacy and creative work.

This research also has some limitations which at the same time provide exciting opportunities for future research. First, the artificial settings of the experiments may affect the generalizability of the results to real-world environments. The controlled experimental conditions do not fully capture the complexities of real-world creative processes where additional factors such as task type and organizational climate[62] can influence creativity. Future research should thus aim to replicate these findings in more naturalistic settings, such as classrooms or professional work environments.

A further methodological limitation of this research is the long-term impact of co-creating (vs. editing) is not tested. Understanding how these roles influence creative self-efficacy and creative outcomes over extended time periods is important[63], as initial improvements may diminish or change with prolonged use. Longitudinal studies are needed to examine the sustained effects of these roles.

Additionally, the expertise level of the human writer was not considered as a potential boundary condition in this research. The benefits of being a co-creator (vs. editor) may vary significantly between novice and expert writers. Experts writers could benefit less from co-creation because they are less likely to take AI-generated text at face value and insufficiently edit this text as a consequence[64]. Future research should investigate how varying levels of expertise affect our reported findings.

Finally, our research does not address whether using a more advanced generative AI system would unearth different results. Although we argue that co-creating (vs. editing) should elicit positive effects for creative self-efficacy and outcomes irrespective of the AI system in question, this assumption does warrant empirical validation. Future research should explore whether our findings hold in other systems, such as GPT-4o or Claude 3.5, to confirm whether the observed benefits of co-creation persist as AI technologies evolve.

Concluding remarks

As the production of creative work is increasingly shaped by generative AI, our conceptualization of what it means be creative has witnessed a notable evolution. We have examined human-AI creative collaboration through two methodologically rigorous experimental studies. Together, our findings suggest that for AI to successfully augment human creativity, it is a requirement that it promotes creative self-efficacy and places humans in the role of a co-creator, not an editor.

Methods

The present research involved no more than minimal risks, and all study participants were 18 years of age or older. All research studies received ethical approval from the Institutional Review Board (IRB) at the National University of Singapore (Ethical Approval Code: BIZ-MNO-20-0213), and all methods of the reported studies were performed in accordance with the ethical guidelines and regulations of this committee (https://www.nus.edu.sg/research/irb/resources/references). Informed consent was obtained for all participants. All manipulations and measures are reported. Data and syntax files are available on the Open Science Framework at https://osf.io/3jqga/.

In Studies 1 and 2, we recruited participants from the online sample recruiting platform Prolific (ProA; http://www.prolific.ac). To be eligible, participant's first language needed to be English. ProA is an online platform explicitly designed for online participant recruitment by the scientific community[65] and has been empirically demonstrated to provide higher quality data than alternative online platforms[66]. Moreover, in both studies, we utilized attention checks, an important method for enhancing data quality[67], particularly when utilizing online platforms such as Prolific[66]. However, only in Study 1 did any participants fail this check (n = 6). Following prior research, we selected participants with an approval rating of at least 97%. Condition assignments were random in both studies, with randomization administered via front-end programming.

In both studies, integrating the poetry generation system with our web application (URL link) required both back- and front-end development. For our back-end framework, we chose Flask[68]—a small and lightweight Python web framework that provides useful tools and features that make creating web applications in Python more convenient. The programming language adopted for this is Python. To record and store data submitted by participants, Firebase was utilized. Firebase is a platform developed by Google that not only facilitates data storage, but also enables the development of mobile and web applications. We opted to use Google Cloud Platform (GCP) as our server and Gunicorn for deployment. Gunicorn is a widely used high-performance Python WSGI HTTP Server based on the UNIX system. For front-end development, we mainly used Jinja2 to implement our functions (e.g., login verification, submission of forms) and facilitate asynchronous transfers (data exchange between back- and front-end). Jinja2 is a full-featured template engine for Python. The interface we developed is built on Bulma—a free, open-source framework that provides ready-to-use front-end components that can be combined to build responsive web interfaces. More complicated components and functions required jQuery and AJAX. The programming languages we adopted for front-end development were HTML, CSS, and JavaScript. HTML is adopted to provide the basic structure of our web application, which in turn is enhanced and modified via the use of CSS and JavaScript. CSS is used to control presentation, formatting, and layout. JavaScript is used to control the execution of specialized functions (e.g., imposing time constraints on webpages).

Ethical approval

All studies research received ethical approval from the Institutional Review Board (IRB) at the National University of Singapore (Ethical Approval Code: BIZ-MNO-20-0213), and all methods of the reported studies were performed in accordance with the ethical guidelines and regulations of this committee (https://www.nus.edu.sg/research/irb/resources/references).

Data availability

Data and syntax code from both studies reported in this paper are publicly available at https://osf.io/3jqga/. Analyses were conducted with SPSS 23.0.

Received: 15 March 2024; Accepted: 5 August 2024
Published online: 09 August 2024

References

1. Lee, M., Liang, P. & Yang, Q. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proc. 2022 CHI Conference on Human Factors in Computing Systems* 1–19 (2022).
2. Jo, A. The promise and peril of generative AI. *Nature* **614**(1), 214–216 (2023).
3. Gozalo-Brizuela, R. & Garrido-Merchan, E. C. ChatGPT is not all you need. A state of the art review of large generative AI models. Preprint at http://arXiv.org/2301.04655 (2023).
4. Mazzone, M. & Elgammal, A. Art, creativity, and the potential of artificial intelligence. *Arts* **8**(1), 26 (2019).
5. Eshraghian, J. K. Human ownership of artificial creativity. *Nat. Mach. Intell.* **2**(3), 157–160 (2020).
6. Anantrasirichai, N. & Bull, D. Artificial intelligence in the creative industries: A review. *Artif. Intell. Rev.* **1**, 1–68 (2022).
7. Girotra, K., Meincke, L., Terwiesch, C. & Ulrich, K. T. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071* (2023).
8. Köbis, N. & Mossink, L. D. Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput. Hum. Behav.* **114**, 106553 (2021).
9. Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K. & Chen, L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *J. Inf. Technol. Case Appl. Res.* **25**(3), 277–304 (2023).
10. De Cremer, D., Bianzino, N. M. & Falk, B. How generative AI could disrupt creative work. *Harvard Bus. Rev.* https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work (2023).
11. Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S. & Wu, C. S. Art or artifice? Large language models and the false promise of creativity. Preprint at http://arXiv.org/2309.14556 (2023).
12. Padmakumar, V. & He, H. Does writing with language models reduce content diversity? Preprint at http://arXiv.org/2309.05196 (2023).
13. Perez, E. *et al*. Discovering language model behaviors with model-written evaluations. Preprint at http://arXiv.org/2212.09251 (2022).
14. Mukherjee, A. & Chang, H. Managing the creative frontier of generative AI: The novelty-usefulness tradeoff. *Calif. Manag. Rev.* **1**, 1 (2023).
15. Doshi, A. R. & Hauser, O. Generative artificial intelligence enhances creativity. *Available at SSRN* (2023).
16. Boussioux, L., N Lane, J., Zhang, M., Jacimovic, V. & Lakhani, K. R. *The Crowdless Future? How Generative AI is Shaping the Future of Human Crowdsourcing. The Crowdless Future. Available at SSRN* (2023).
17. Amabile, T. M. Social psychology of creativity: A consensual assessment technique. *J. Person. Soc. Psychol.* **43**, 5 (1982).
18. Baer, J. The importance of domain-specific expertise in creativity. *Roeper. Rev.* **37**(3), 165–178 (2015).
19. Kenny, D. A. Enhancing validity in psychological research. *Am. Psychol.* **74**(9), 1018 (2019).
20. Aiyappa, R., An, J., Kwak, H., & Ahn, Y. Y. (2023). Can we trust the evaluation on ChatGPT? Preprint at http://arXiv.org/2303.12767.
21. Chen, L., Zaharia, M. & Zou, J. How is ChatGPT's behavior changing over time? Preprint at http://arXiv.org/2307.09009 (2023).
22. Tierney, P. & Farmer, S. M. Creative self-efficacy development and creative performance over time. *J. Appl. Psychol.* **96**(2), 277 (2011).
23. Noy, S. & Zhang, W. *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. Available at SSRN 4375283* (2023).
24. Mirbabaie, M., Brünker, F., Möllmann, N. R. & Stieglitz, S. The rise of artificial intelligence–understanding the AI identity threat at the workplace. *Electron. Mark.* **1**, 1–27 (2022).

25. Tang, P. M. *et al.* When conscientious employees meet intelligent machines: An integrative approach inspired by complementarity theory and role theory. *Acad. Manag. J.* **65**(3), 1019–1054 (2022).
26. Roskes, M. Constraints that help or hinder creative performance: A motivational approach. *Creat. Innov. Manag.* **24**(2), 197–206 (2015).
27. Dell'Acqua, F. *et al.* Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper, (24–013)* (2023).
28. Tierney, P. & Farmer, S. M. Creative self-efficacy: Its potential antecedents and relationship to creative performance. *Acad. Manag. J.* **45**(6), 1137–1148 (2002).
29. McCormack, J. *et al.* Design considerations for real-time collaboration with creative artificial intelligence. *Org. Sound* **25**(1), 41–52 (2020).
30. Bandura, A. & Schunk, D. H. Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *J. Person. Soc. Psychol.* **41**(3), 586 (1981).
31. Shelley, P. B. & Brett-Smith, H. F. B. *A defence of Poetry* 79 (Brodie, 1980).
32. Perkins, D. *A history of Modern Poetry* Vol. 1 (Harvard University Press, 1976).
33. Van de Cruys, T. Automatic poetry generation from prosaic text. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 2471–2480 (2020).
34. Lee, D. & Seung, H. S. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.* **13**, 1 (2000).
35. Amabile, T. M. Motivation and creativity: Effects of motivational orientation on creative writers. *J. Person. Soc. Psychol.* **48**(2), 393 (1985).
36. Amabile, T. M., Hennessey, B. A. & Grossman, B. S. Social influences on creativity: The effects of contracted-for reward. *J. Person. Soc. Psychol.* **50**(1), 14 (1986).
37. Kaufman, J. C., Baer, J. & Cole, J. C. Expertise, domains, and the consensual assessment technique. *J. Creat. Behav.* **43**(4), 223–233 (2009).
38. Hayes, A. F. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (Guilford Publications, 2017).
39. Carroll, J. M. Human–computer interaction: Psychology as a science of design. *Annu. Rev. Psychol.* **48**(1), 61–83 (1997).
40. Xu, W. Toward human-centered AI: A perspective from human–computer interaction. *Interactions* **26**(4), 42–46 (2019).
41. Bai, H. X. *et al.* Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology* **296**(3), E156–E165 (2020).
42. Lebovitz, S., Lifshitz-Assaf, H. & Levina, N. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Org. Sci.* **33**(1), 126–148 (2022).
43. Marr, B. *Boost Your Productivity with Generative AI. Harvard Business Review*. https://hbr.org/2023/06/boost-your-productivity-with-generative-ai (2023).
44. Hennessey, B. A. Self-determination theory and the social psychology of creativity. *Psychol. Inq.* **11**(4), 293–298 (2000).
45. Sheldon, K. M. Creativity and self-determination in personality. *Creat. Res. J* **8**(1), 25–36 (1995).
46. Jachimowicz, J. M., Duncan, S., Weber, E. U. & Johnson, E. J. When and why defaults influence decisions: A meta-analysis of default effects. *Behav. Public Policy* **3**(2), 159–186 (2019).
47. Biddle, B. J. Recent developments in role theory. *Annu. Rev. Sociol.* **12**(1), 67–92 (1986).
48. Eagly, A. H. & Wood, W. Social role theory. *Handb. Theor. Soc. Psychol.* **2**, 458–476 (2012).
49. Jiao, C., Richards, D. A. & Hackett, R. D. Organizational citizenship behavior and role breadth: A meta-analytic and cross-cultural analysis. *Hum. Resour. Manag.* **52**(5), 697–714 (2013).
50. Farmer, S. M., Tierney, P. & Kung-McIntyre, K. Employee creativity in Taiwan: An application of role identity theory. *Acad. Manag. J.* **46**(5), 618–630 (2003).
51. McAllister, D. J., Kamdar, D., Morrison, E. W. & Turban, D. B. Disentangling role perceptions: How perceived role breadth, discretion, instrumentality, and efficacy relate to helping and taking charge. *J. Appl. Psychol.* **92**(5), 1200 (2007).
52. Mednick, S. The associative basis of the creative process. *Psychol. Rev.* **69**(3), 220 (1962).
53. Mehta, R. & Zhu, R. Blue or red? Exploring the effect of color on cognitive task performances. *Science* **323**(5918), 1226–1229 (2009).
54. Nijstad, B. A. & Stroebe, W. How the group affects the mind: A cognitive model of idea generation in groups. *Person. Soc. Psychol. Rev.* **10**(3), 186–213 (2006).
55. Moraru, A., Memmert, D. & van der Kamp, J. Motor creativity: The roles of attention breadth and working memory in a divergent doing task. *J. Cogn. Psychol.* **28**(7), 856–867 (2016).
56. Simon, H. A. A behavioral model of rational choice. *Q. J. Econ.* **69**, 99–118 (1955).
57. Brucks, M. S. & Levav, J. Virtual communication curbs creative idea generation. *Nature* **605**(7908), 108–112 (2022).
58. Anantrasirichai, N. & Bull, D. Artificial intelligence in the creative industries: A review. *Artif. Intell. Rev.* **55**(1), 589–656 (2022).
59. McGuire, J. *et al.* The reputational and ethical consequences of deceptive chatbot use. *Sci. Rep.* **13**(1), 16246 (2023).
60. Lai, M. C. & Chen, Y. C. Self-efficacy, effort, job performance, job satisfaction, and turnover intention: The effect of personal characteristics on organization performance. *Int. J. Innov. Manag. Technol.* **3**(4), 387 (2012).
61. Chen, L., Chen, P. & Lin, Z. Artificial intelligence in education: A review. *IEEE Access* **8**, 75264–75278 (2020).
62. Mutonyi, B. R., Slåtten, T. & Lien, G. Organizational climate and creative performance in the public sector. *Eur. Bus. Rev.* **32**(4), 615–631 (2020).
63. Putz, L. M., Hofbauer, F. & Treiblmaier, H. Can gamification help to improve education? Findings from a longitudinal study. *Comput. Hum. Behav.* **110**, 106392 (2020).
64. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In *Proc. ACM on Human-computer Interaction, 5(CSCW1)* 1–21 (2021).
65. Palan, S. & Schitter, C. Prolific.ac—A subject pool for online experiments. *J. Behav. Exp. Financ.* **17**, 22–27 (2018).
66. Peer, E., Brandimarte, L., Samat, S. & Acquisti, A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *J. Exp. Soc. Psychol.* **70**, 153–163 (2017).
67. Oppenheimer, D. M., Meyvis, T. & Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* **45**(4), 867–872 (2009).
68. Grinberg, M. *Flask Web Development: Developing Web Applications with Python* (O'Reilly Media Inc, 2018).

Author contributions

J.M. and D.D.C. designed the studies; T.V.D.C. and J.M. programmed the experiments. J.M. collected and analysed the data; J.M. wrote a first draft of the manuscript, all the other authors subsequently revised the manuscript.

Funding

The funding was provided by Ministry of Education—Singapore (11/2019/115).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024