RESEARCH-ARTICLE

How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment

JOSHUA ASHKINAZE, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

JULIA MENDELSOHN, University of Maryland, College Park, College Park, MD, United States

LI QIWEI, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

CEREN BUDAK, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

ERIC E GILBERT, University of Michigan, Ann Arbor, Ann Arbor, MI, United States

How AI Ideas Affect the Creativity, Diversity, and Evolution of Human Ideas: Evidence From a Large, Dynamic Experiment

Joshua Ashkinaze
University of Michigan
Ann Arbor, USA
jashkina@umich.edu

Julia Mendelsohn
University of Maryland
Maryland, USA
juliame@umd.edu

Li Qiwei
University of Michigan
Ann Arbor, USA
rrll@umich.edu

Ceren Budak
University of Michigan
Ann Arbor, USA
cbudak@umich.edu

Eric Gilbert
University of Michigan
Ann Arbor, USA
eegg@umich.edu

Abstract

Exposure to large language model output is rapidly increasing. How will seeing AI-generated ideas affect human ideas? We conducted a dynamic experiment (800+ participants, 40+ countries) where participants viewed creative ideas that were from ChatGPT or prior experimental participants, and then brainstormed their own idea. We varied the number of AI-generated examples (none, low, or high *exposure*) and if the examples were labeled as "AI" (*disclosure*). We find that high AI exposure (but not low AI exposure) did not affect the creativity of individual ideas but did increase the average amount and rate of change of collective idea diversity. AI made ideas different, not better. There were no main effects of disclosure. We also found that self-reported creative people were less influenced by knowing an idea was from AI and that participants may knowingly adopt AI ideas when the task is difficult. Our findings suggest that introducing AI ideas may increase collective diversity but not individual creativity.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**; **Collaborative interaction**; • **Computing methodologies** → **Artificial intelligence**; **Natural language processing**;

Keywords

artificial intelligence, large language models, human-computer interaction, creativity, collective intelligence, cultural evolution
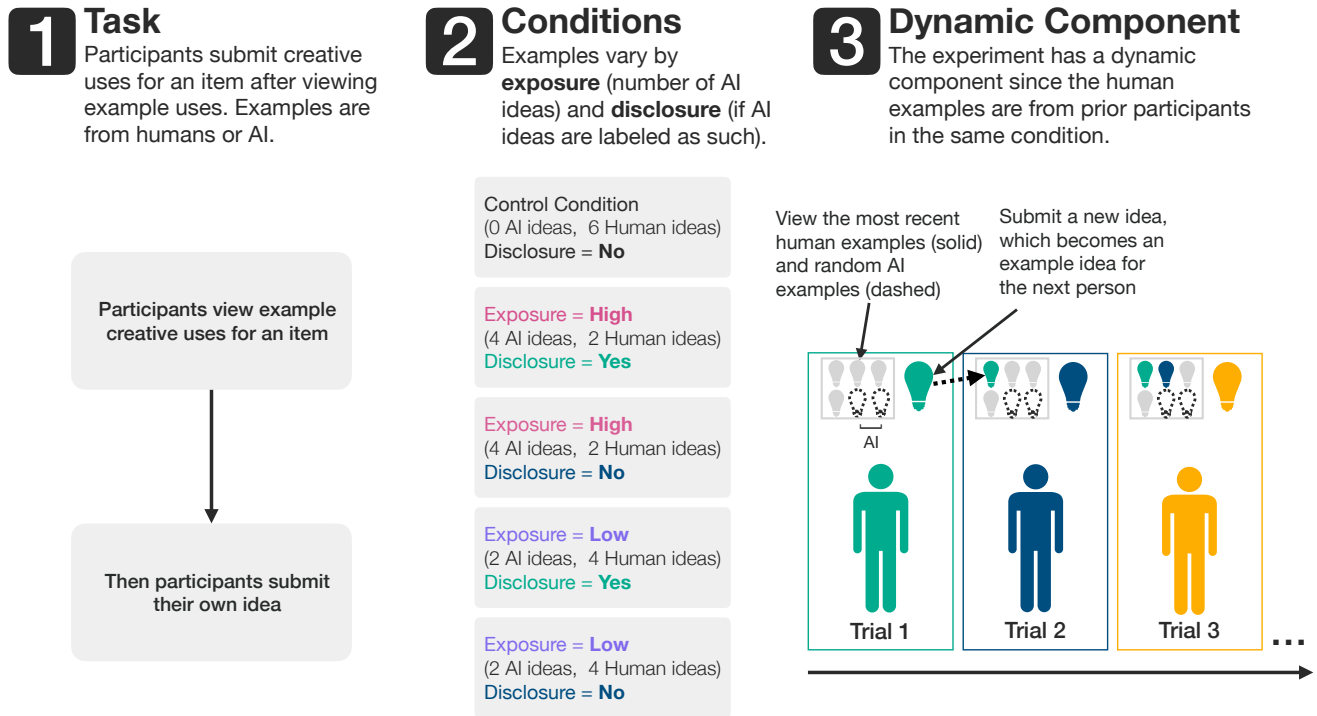
1 Introduction

If we think of culture as a feedback loop where individuals and societies shape each other through exchanges of ideas and practices [10, 65], then what happens when generative AI enters this "culture loop"? Exposure to LLMs (large language models) is increasing: When released, ChatGPT was the fastest-growing consumer application in history [34]. Moreover, we are likely exposed to even more AI content than we realize: Humans overestimate their ability to distinguish AI from human content [37]. This exposure likely matters: Ideas we see affect the ideas we create [52]. How, then, will the rapid rise of exposure to LLM-generated ideas affect the creativity, diversity, and evolution of human ideas? And to what extent *do* AI ideas influence human ideas? These questions have implications for how AI may shape collective intelligence [14, 17, 60]. Our study speaks to these questions in the specific context of brainstorming, an activity that occurs across many domains.

The widespread dissemination of AI-generated content creates a phenomenon we call "passive exposure," which differs from the "active engagement" primarily studied in prior human-AI interactions. By "passive exposure", we refer to cases when (A) users see LLM outputs but do not have an active role in the creation of these outputs and (B) users are given no instructions to actively engage with these outputs. Passive exposure approximates how users often encounter LLM outputs in the real world. For example, OpenAI users generate 100 billion words per day [28]. It is likely that the number of people who are merely seeing (i.e., *passively exposed*) AI output is significantly larger than the number of people who are creating (i.e., *actively engaging*) with these systems. Yet in existing studies of human-AI creativity, participants are often *actively* interacting with an AI system [11, 25, 41, 57, 58, 78].

As AI exposure has increased, so have concerns over AI *disclosure* [32] (whether providers should disclose when they use AI systems). California considered requiring disclosure on behalf of anyone using bots on social media [77]. Concerns regarding the disclosure of LLMs are only more likely to grow. LLM output is increasingly indistinguishable [37] from that of humans. We are interested if disclosing ideas as coming from AI moderates the effect of AI exposure.

Motivated by these dynamics, we conducted a large-scale experiment to systematically examine how AI exposure and disclosure influence the creativity, diversity, and evolution of human ideas. We utilize a variant of the Alternate Uses Task (AUT, [30]), a standard

1 Task

Participants submit creative uses for an item after viewing example uses. Examples are from humans or AI.

Participants view example creative uses for an item

↓

Then participants submit their own idea

2 Conditions

Examples vary by **exposure** (number of AI ideas) and **disclosure** (if AI ideas are labeled as such).

Control Condition
(0 AI ideas, 6 Human ideas)
Disclosure = **No**

Exposure = **High**
(4 AI ideas, 2 Human ideas)
Disclosure = **Yes**

Exposure = **High**
(4 AI ideas, 2 Human ideas)
Disclosure = **No**

Exposure = **Low**
(2 AI ideas, 4 Human ideas)
Disclosure = **Yes**

Exposure = **Low**
(2 AI ideas, 4 Human ideas)
Disclosure = **No**

3 Dynamic Component

The experiment has a dynamic component since the human examples are from prior participants in the same condition.

View the most recent human examples (solid) and random AI examples (dashed)

Submit a new idea, which becomes an example idea for the next person

AI

Trial 1   Trial 2   Trial 3   ...

Figure 1: Graphical depiction of experiment. The task (Panel 1) is to submit a creative idea after seeing examples, where examples are from humans or AI. We vary (Panel 2) the amount of AI ideas in the example set (*exposure*) and if AI ideas are labeled as such (*disclosure*). The experiment is dynamic (Panel 3). Responses from prior participants serve as examples for future participants.

measure of creativity, and manipulate exposure to LLM-generated ideas. In the AUT, participants generate non-obvious uses for a common item. For example: *What is a creative use for a tire?* In our variant, participants complete the AUT after viewing example ideas, which serve as our experimental manipulation. Examples vary in AI exposure (none, few, or many AI examples) and AI disclosure (whether AI-generated ideas are explicitly labeled). The human-generated ideas in each example set originate from prior participants in the same experimental condition. See Figure 1 for a depiction.

Our dynamic experiment design leverages ideas from prior participants as stimuli for future participants, reflecting the interdependent nature of brainstorming. Ideas build upon previous ideas. This approach captures (certain aspects of) the compounding effects of integrating LLMs into the "culture loop" and simulates potential futures for human-AI collaboration. Our design enables us to observe both average performance levels and temporal dynamics of creativity and diversity across conditions. This setup provides insights into how LLMs shape collective thought.

1.1 Findings

Our experiment manipulates two key factors: (1) AI exposure (the number of AI-generated ideas in example sets) and (2) AI disclosure (whether AI ideas are labeled as such). We investigate three

outcomes: individual creativity, idea diversity (measured at individual and collective levels), and AI idea adoption (who and when adopts AI example ideas). For creativity and diversity, we analyze both average levels and temporal evolution across experimental iterations. See Table 1 for our measures. Here are our findings.

(1) **High AI exposure increases collective diversity but not individual creativity.** We find that AI exposure did not affect individual creativity. However, conditions with high levels of AI exposure had more *collective* idea diversity. That is, ideas in the high AI exposure conditions were more *different* from each other but not necessarily better. Our findings around creativity and diversity suggest the effect of AI exposure may be nuanced.

(2) **High AI exposure increases the speed at which idea diversity develops.** Through our dynamic design, we find that high AI exposure increases not only the average levels of collective idea diversity but also the rate of change in idea diversity. This is a consequential finding since even small differences in rates of change can lead to large cumulative differences over time.

(3) **People who identify as creative are less influenced by AI disclosure**. We find that for users who self-identify as highly creative, adoption of AI ideas is not influenced by AI disclosure. But AI disclosure did affect the adoption of AI

ideas for users who self-identified as low in creativity. This suggests that highly creative people will not be "duped" into adopting AI ideas.

(4) **Participants may adopt AI ideas for harder prompts.** We find that when AI ideas are disclosed as such, participants were more likely to adopt the ideas of AI for difficult AUT prompts. This suggests that users may rely on AI ideas not for trivial creative tasks but for difficult ones. Since this finding is based on a small number of AUT items, this is speculative.

1.2 Defining Concepts and Variables

1.2.1 Creativity. Creativity is defined in many ways [75]. But one common conception is divergent thinking [29]: When "an individual solves a problem or reaches a decision using strategies that deviate from commonly used or previously taught strategies" [53]. One of the most common [1] tests of divergent thinking is the Alternate Uses Task (AUT) [30][1], where participants are asked to think of an original use for an everyday object. Traditionally, responses to the AUT are measured along four dimensions: originality (how original the idea is), elaboration (how much the participant elaborates on the idea), fluency (how many ideas), and flexibility (different categories of ideas). The latter two can only be measured if the participant provides multiple responses to the same question. Due to our research design[2], we have participants generate just one creative idea (as in [6]), and we focus on originality.

We follow a tradition of scoring responses to the AUT computationally [5, 6, 21, 56, 79, 82]. Specifically, we measure the **creativity** of AUT ideas with an existing fine-tuned GPT-3 classifier [55], which has an r=0.81 overall correlation with human judgments of AUT originality. Moreover, we chose AUT items for our experiment where the classifier had the *highest* accuracy[3]. Our task is highly 'in-domain' for the classifier: we ask participants to do the *same exact task* for the *same exact items* the model was trained on. We refer to the originality score from this classifier as individual-level **creativity**, but future work can explore other dimensions of creativity (such as fluency). We discuss this classifier more in Appendix A.1.

1.2.2 Idea Diversity & AI Adoption. In addition to creativity, we measure how our experimental factors (LLM exposure and LLM disclosure) shape the *diversity* of ideas that participants produce. This is a complementary measure to creativity. Creativity is often thought of as an individual outcome. Diversity is a collective outcome. Equivalently, creativity is a property of an *idea* while diversity is a property of an *idea set.* We measure two sides of diversity—semantic *divergence* (which we refer to as **idea diversity**) and semantic *convergence* towards AI ideas (which we refer to as **AI adoption**).

To measure idea diversity and AI adoption, we first embed all ideas using SBERT [62], which are transformer-based embeddings

designed for sentences. SBERT excels at capturing semantic similarity [62]. Prior work uses neural embeddings to compute similarity for AUT responses [4] and other creative tasks [67].

> **Idea diversity** is the median pairwise cosine distance between idea embeddings in an idea set. As robustness checks, we also measure the mean pairwise distance and average distance to the centroid of a set.
>
> **AI adoption** is the maximum cosine similarity between the embedding of the idea a participant submits and the embeddings of AI examples that the participants see. Following [67], we use the max rather than a measure of central tendency because if a participant is inspired by an idea, it would likely be a *single* idea. As robustness checks, we also measure the mean and median pairwise similarity between the submitted idea and an AI example, but these are noisier measures of adoption.

2 Related Work

Our work bridges three research streams: human-AI co-creation, crowd-sourced creativity, and collective dynamics. AI ideas are scattered amongst human ideas, whether or not we can tell [37]. This exposure presumably affects the ideas we create (co-creation). And our ideas presumably affect the ideas *others* create (crowdsourced creativity). While real-world culture is dynamic and evolving, most experiments are not set up to capture evolution (collective dynamics). By employing a large-N sample size and "multiple-worlds" setup, we model the complex dynamics of AI influence. In Appendix C, we provide a more detailed overview of how our experimental factors (LLM exposure and disclosure) may affect creativity and diversity of ideas, as well as how attitudinal variables may moderate these effects. Overall, predictions from prior work and theory are conflicting—motivating the present study.

2.1 Human-AI Co-Creation

As AI becomes more creative [43], researchers have explored how co-creating with AI affects human creativity. Much of this research explores creative writing with language models [19, 24–26, 33, 41, 44, 45, 58, 67, 78, 83]. While most prior work involves users *actively engaging* with custom systems, our study is concerned with *passive exposure* to off-the-shelf models. The relationship between AI ideas and their effect on human creativity is nuanced. Task-level and attitudinal factors play a role. For example, seeing AI examples influenced outcomes for hard, but not easy, prompts in Roemmele [67], and Gero et al. [26] found the quality of LLM outputs did not correlate with perceived usefulness. Large variance in the perceived usefulness of outputs from co-creation systems [15] suggests human attitudes partially determine the utility of AI creativity aids. We extend this predominantly qualitative work with a large-scale quantitative study.

Several studies, like ours, explored how (post-ChatGPT) generative artificial intelligence affects creativity and diversity. Some studies show that ideating with generative AI decreases diversity [2, 20, 58]. Other studies suggest generative AI increases individual creativity [18, 20] while others find mixed effects on creativity [2]. Our study offers several additions. First, because of its dynamic design, we test the long-run effects of AI, where ideas feed forward

---

[1]https://www.mindgarden.com/67-alternate-uses

[2]Participants see the most recent responses in the condition as stimuli, so if one participant brainstorms many responses, that participant would be over-represented in future participants' example sets.

[3]*tire* (r=0.91), *pants* (r=0.91), *shoe* (r=0.91), *table* (r=0.9), and *bottle* (r=0.88)

to future participants. Second, our study is concerned with "passive exposure": Participants are not told ideas are from AI, and are not instructed to engage with these ideas. By systematically ablating whether AI ideas are disclosed as such, we can explore if the effect of AI ideas depends on knowledge of where the idea is from. Third, we employ a large sample size—which is useful since it provides power for precise estimates of effects and the ability to capture heterogeneity. Fourth, our large sample is comprised of creative professionals and technology-oriented users, two groups relevant to the phenomenon.

2.1.1 Crowdsourced Creativity. Crowdsourcing can enhance creative outcomes [35, 51, 72, 80, 81]. Siangliulue et al. [72] found that the creativity and diversity of idea sets that participants saw influenced the creativity and diversity of what these participants produced. This supports a main contention of our paper. AI exposure matters because the ideas we see affect the ideas we create. We incorporate elements of crowdsourced creativity, particularly in measuring how creativity and diversity unfold over subsequent generations.

2.1.2 Collective Dynamics & Multiple-Worlds Experiments. Prior work in complex systems and sociology highlights the importance of studying *collectives* to understand social dynamics [69]. Meanwhile, traditional experiments focus on *individuals*. Identifying the effect of AI ideas on the diversity and evolution human ideas similarly requires an examination of complex systems as opposed to individuals in isolation. Hence, our experimental design draws on the "multiple-world" paradigm (e.g., [68, 69]): We create multiple, parallel realizations of worlds with and without AI ideas, each evolving independently under controlled conditions. By employing a large-N sample size and many different parallel worlds, we can better understand the collective effects of AI ideas on human ideas. Beyond simple averages, we can also model how AI ideas affect the *evolution* of human ideas.

2.1.3 Our Contributions. Our study integrates human-AI co-creation, crowd creativity, and collective dynamics. Traditional co-creation studies typically confound exposure and disclosure effects: when creating with AI, one cannot separate the AI's *content* from the *knowledge* that this content is AI-generated. Our design allows us to estimate these effects independently. Co-creation studies typically employ a small number of specialized participants actively engaged with a system. We are interested in the effects of (1) *passive exposure* on (2) a more general public. For this reason, we adopt a large-scale experimental design—similar to crowd-sourced creativity studies—that lets us estimate effects on the general public rather than specialized users. A key benefit of our large sample size is that we can precisely estimate how participant attitudes affect human-AI outcomes. This is important because attitudes affect human-AI collaboration outcomes [26]. Drawing on the "multiple-worlds" paradigm [69], our experiment design also lets us understand the effect of AI *over time* since responses feed forward, allowing us to observe differences in *rates of change* between conditions. This method can be used for other experiments to measure the cumulative effect of AI.

3 Experiment

3.1 Summary

We recruited participants through social media platforms and newsletters. Once participants clicked on the link to the experiment, they were taken to a landing page. In addition to a consent button, that landing page asked several questions related to their self-perceived creativity, attitudes towards AI, and demographics. Following consent, participants completed 5 trials. Each trial required generating a creative use for an object under a specific experimental condition after viewing example ideas—which were our experimental manipulation. Ideas fed forward to future trials such that if a participant was in the {[*Control*], *tire*} condition the example ideas the participant saw were the most recent ideas from *prior* participants in the {[*Control*], *tire*} condition. See Figure 1 for experimental conditions. The experiment was approved by our University's IRB.

3.2 Stimuli Creation

Picking AUT Items. Previously, Organisciak et al. [56] fine-tuned GPT-3 to predict the creativity of AUT items. Since we use this classifier for our creativity DV, we picked the five items for which the classifier had the highest classifier-human judgment correlations: *tire* (r=0.91), *pants* (r=0.91), *shoe* (r=0.91), *table* (r=0.9), and *bottle* (r=0.88). Our task is highly 'in-domain' for the classifier: Participants do the *same task*, using the *same items*, from the classifier training and evaluation. See Appendix A.1 for more details on the classifier dataset ("Organisciak Dataset"), training, and performance.

Model & Prompting Strategy. We generated AI ideas using ChatGPT-3.5 with a zero-shot prompt adapted from [73]. We chose ChatGPT-3.5 for ecological validity as it was the freely available model powering ChatGPT at the time of our study. Zero-shot prompting required no labeled data, making it more practical for typical users. We modified the prompt from Stevenson et al. [73] to match word counts from the Organisciak Dataset [56], preventing easy identification of AI-generated content. Using parameters from [73] and instructions to match human word counts, our modified prompt produced responses similar to human responses in length. These generated ideas served as stimuli for our main experiment. For prompt details and statistical tests, see Appendix A.2.

3.3 Recruitment Strategy

We recruited volunteer participants from three sources: (1) Facebook ads, (2) Reddit, and (3) Creative Mornings' weekly newsletter. Creative Mornings[4], "the world's largest face-to-face creative community," is an organization for creative professionals that hosts talks and meetups. Creative Mornings agreed to distribute our experiment. We verified all participants were at least 18 years old. While we did not offer monetary compensation, we offered to give participants information about themselves, such as their creativity relative to both humans and AI, and their ability to spot creative ideas. Providing information to participants about themselves is often effective for recruiting volunteers since it makes the task intrinsically rewarding [63]. See Table 3 for sources and SI 6 for sample characteristics.

---

[4]https://www.creativemornings.com/
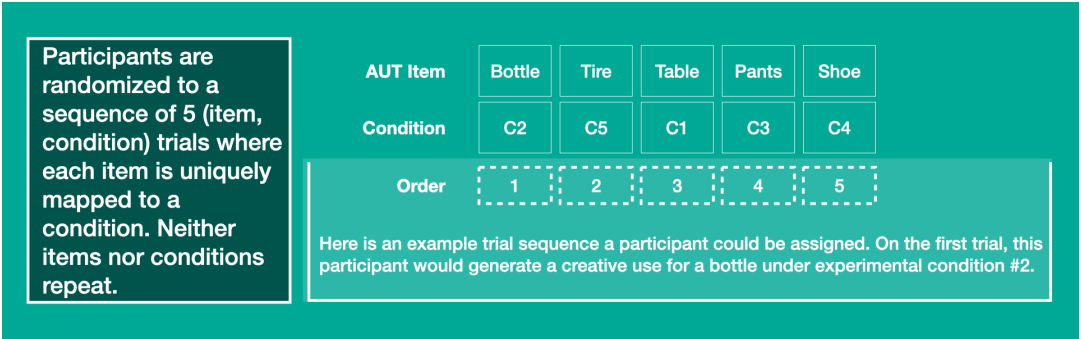
Figure 2: Participants are randomized to a sequence of 5 trials. In each trial, participants generate a creative use for an item under a specific experimental condition. Neither items nor conditions repeat in a 5-trial sequence.

We recruited volunteer participants instead of crowdsourced workers for several reasons. First, we wanted participants to be intrinsically motivated since (1) many theories suggest intrinsic motivation helps creativity [47] and (2) we did not want low-quality engagement to confound results (especially since ideas propagate forward). Second, we recruited participants in a targeted manner. In particular, we wanted to target (1) individuals who have a demonstrated interest in technology and (2) those who have a demonstrated interest in creativity. These groups are most relevant to the phenomenon in question. We reached technology-oriented users by posting the experiment in the following subreddits: r/chatgpt, r/InternetIsBeautiful, r/singularity, and r/artificial. We reached creativity-oriented users by posting the experiment in r/writing, r/poetry, and the Creative Mornings newsletter. We also used several 'neutral' sources to test the experiment: r/samplesize and Facebook ads. If a participant completed the experiment, then the participant was given a shareable link to their results so they could spread the study.

3.4 Experiment Procedure

Once participants clicked on our link, they went to a landing page that included a consent form, task description, and pre-treatment questions. See Figure 3 for the task description. See SI 5 for more details on how participant feedback was calculated.

3.4.1 Pre-Treatment Questions. Participants were asked several pre-treatment questions (Figure 4), including self-perceived creativity, perceived creativity of AI, and attitude towards AI. The third question was from Pew [48, 49]. We chose the Pew question instead of a longer battery of questions about AI to minimize the response burden. See SI 2 for more details about these questions.

3.4.2 Randomization. Participants were assigned a sequence of 5 trials, where each trial was a {[condition], item} pair. For example, one trial might be a creative idea for *pants* in the [*High Exposure, Disclosed*] condition. We mapped each AUT item (pants, tire, shoe, bottle, table) to one of the five conditions such that neither conditions nor items repeated in a 5-item sequence. See Figure 2 for a visual explanation.

3.4.3 Task Instructions. For each trial (Figure 5), we asked participants to first rank a list of example ideas by creativity and then

What you will do:
We'll show you 5 common items, and you'll come up with creative uses for each item. To spark your imagination, you'll see ideas from prior participants and even from AI (i.e., ChatGPT). You'll be asked to rank these ideas in order of creativity. The ideas you write may be anonymously shown to future participants to spark their imagination. The study takes 3-6 minutes to complete. The goal is to learn about how humans and AI brainstorm.

What you will learn:
- How creative you are compared to other humans
- How creative you are compared to AI
- How well you can rank creative ideas

We will give you a shareable link with results at the end.

Figure 3: Study description provided to participants.

submit their own idea. See Appendix D for screenshots. We included the ranking task to ensure engagement with the example ideas. Across conditions, we manipulated two factors: (1) *exposure*—whether AI-generated ideas were included in the example set, and (2) *disclosure*—whether AI ideas were labeled. We used the same prompt for human participants as Stevenson et al. [73] used with ChatGPT, with a minor modification requesting a single idea (for reasons discussed in 1.2.1). This prompt incorporates language aligned with divergent thinking assessment best practices [7]. After idea submission, participants received feedback on their idea's uniqueness and their ranking accuracy (screenshots in Appendix D).

3.4.4 Response Chains.

Logic. The human ideas that participants saw came from prior participants in the same {[condition], item} combination. See Figure
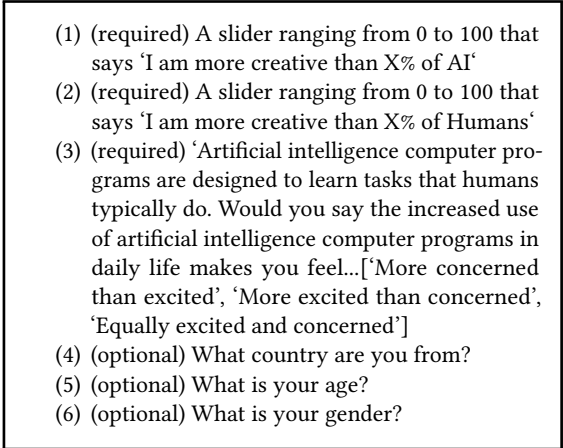
(1) (required) A slider ranging from 0 to 100 that says 'I am more creative than X% of AI'
(2) (required) A slider ranging from 0 to 100 that says 'I am more creative than X% of Humans'
(3) (required) 'Artificial intelligence computer programs are designed to learn tasks that humans typically do. Would you say the increased use of artificial intelligence computer programs in daily life makes you feel...['More concerned than excited', 'More excited than concerned', 'Equally excited and concerned']
(4) (optional) What country are you from?
(5) (optional) What is your age?
(6) (optional) What is your gender?

Figure 4: Pre-Treatment Questions

Task

For this task, you will submit a creative use for a [ITEM]. But before submitting your idea, here are some ideas for inspiration. Rank them by creativity.

Rank Previous Ideas

- Rank these ideas in order of creativity, with the most creative use on top. Drag ideas to rank them.
- We'll show you how your rankings compare to rankings from a highly accurate model.

[SORTABLE EXAMPLE IDEAS HERE]

Submit Your Idea

Your turn! What is a creative use for a [ITEM]? The goal is to come up with a creative idea, which is an idea that strikes people as clever, unusual, interesting, uncommon, humorous, innovative, or different. List a creative use for a [ITEM].
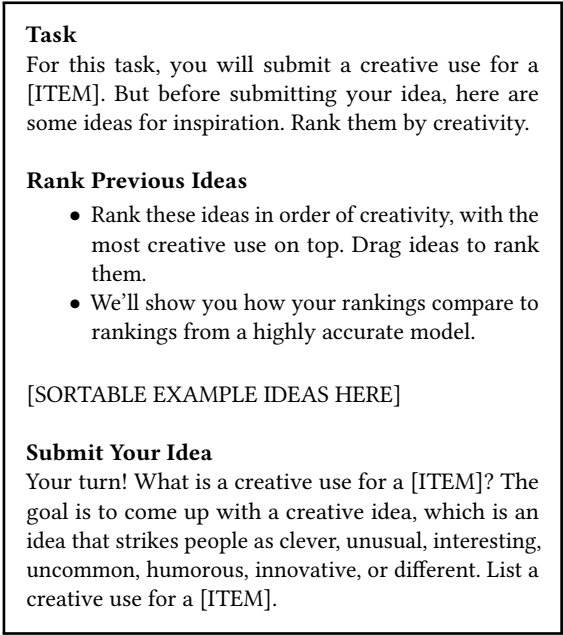
Figure 5: Task Instructions

6. For instance, if a user was placed in the [*Control*] condition for a *tire*, that user would see six human ideas—the most recent six ideas for a *tire* under the [*Control*] condition. In order to avoid overfitting to a specific idea sequence, we reset this 'response chain' every 20 trials. So, the first 20 participants in the {[*Control*], *tire*} combination would see each other's ideas, but the chain would reset for the 21st respondent. We use the logic described in this paragraph and Figure 6 for the human ideas in all conditions.

Note that because human ideas are propagated at the {[*condition*], *item*} level, the human ideas in the [*Control*] condition are 'clean' from AI contamination. They were brainstormed after seeing sets of human-only ideas, also from the [*Control*] condition.

Response Chain Logic

Responses feed forward to future participants. This is an example for the control condition but we do this for all conditions.
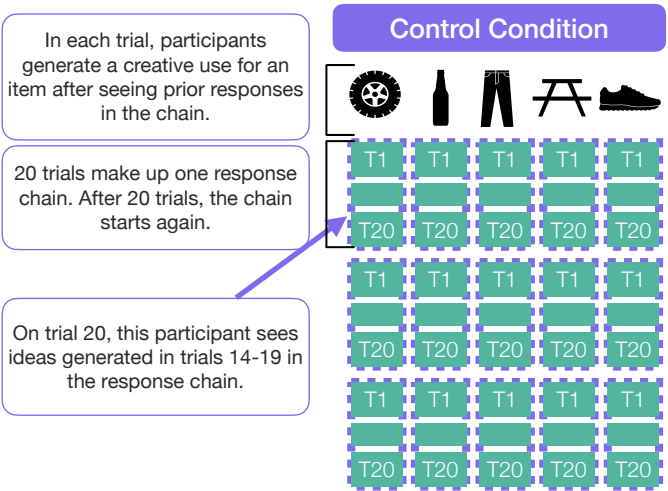

Figure 6: Participants see example ideas from prior participants in the same condition. These 'response chains' reset every 20 responses.

We ran seven response chains for each of the 25 (5 items × 5 conditions) combinations, corresponding to 175 response chains in all and 3500 targeted responses (175 response chains × 20 trials per chain).

Human Seeds. Of course, there is a bootstrapping problem—what human ideas does the *first* person in the {[*Control*], *tire*} condition see? The seeds for each {[*condition*], *item*} combination came from prior responses from the Organisciak Dataset. That is, Participant 1 for a {[*Control*], *tire*} response chain would see 6 seed items. Then Participant 2 in the same response chain would see 5 seed items plus Participant 1's idea (the order of ideas is randomized). Participant 3 would see 4 seed items plus Participant 1 and Participant 2's ideas, etc. We chose a random sample of seeds for each {[*condition*], *item*} combination from the Organisciak Dataset. The dataset labeled ideas with gold-standard human ratings of originality. We conducted an ANOVA and found no significant condition-level difference in the originality of the seeds we used.

4 Outcome Measures

We have three outcome measures (idea diversity, creativity, and AI adoption) and three levels of analysis (local, evolution, and global). See Table 1. The *local* level measures outcomes at the level of an individual trial (e.g., how a submitted response relates to example responses). The *evolution* level measures the rate of change of outcome variables with respect to the trial number in the response chain (i.e., experiment iteration). The *global* level compares all submitted responses in a condition to each other. For all pairwise comparisons,

Table 1: Three outcome measures (idea diversity, creativity, AI adoption) across three levels of analysis (local, global, evolution). All ideas embedded using SBERT.

	Local	Global	Evolution
Creativity	How creative is the submitted response? Measured by the prediction from the Organisciak et al. classifier.	Not applicable	Does the creativity of submitted ideas change over time? Measured by the slope of trial number (i.e., iteration in response chain) on creativity scores from Organisciak et al.
Idea Diversity	How different is a participant's response from example responses? Measured by median pairwise semantic distance between ideas a participant sees and their response (SI 8.3).	How diverse were all participants' ideas in a condition? Measured by median pairwise distance between all submitted ideas in a condition.	Do ideas become more different from each other as the experiment progresses? We measure the median pairwise distance of submitted ideas at each trial number, then calculate the slope of trial number on idea diversity.
AI Adoption	How similar is a participant's response to AI example responses? Measured by the maximum semantic distance between a participant's response and AI examples.	Not applicable	Not applicable

we use a Holm-Bonferroni adjustment for multiple comparisons. For idea diversity and AI adoption, we scale the dependent variable (cosine distance or cosine similarity, respectively) by 100 for easier interpretation. See Appendix B for more methodological details.

5 Results

5.1 Recruited Participants

We received over 3,000 responses from 48 countries. See SI 6 for sample characteristics. Out of a total of five trials, participants finished four trials on average (Table 2), suggesting the experiment was engaging. Most participants came from the Creative Mornings newsletter or r/InternetIsBeautiful (Table 3 for source counts and categorization). Of the non-missing responses, the sample was 50% women, 43% men, 4% non-binary, 3% not disclosed, 1% self-described. The mean age was 34.92 (SD = 10.86). Regarding AI, the sample was 48% neutral, 28% excited, 24% concerned. Participants said they were more creative than 57.86% (SD = 26.66) of AI and 58.67% (SD = 23.65) of humans. See SI 6 for kernel density plots.

Table 2: Summary Statistics of Experiment

	Value
Unique Countries	48
Total Responses	3414
Unique Participants	844
Avg Responses/Participant	4.05
Avg Duration/Response (Seconds)	144.31
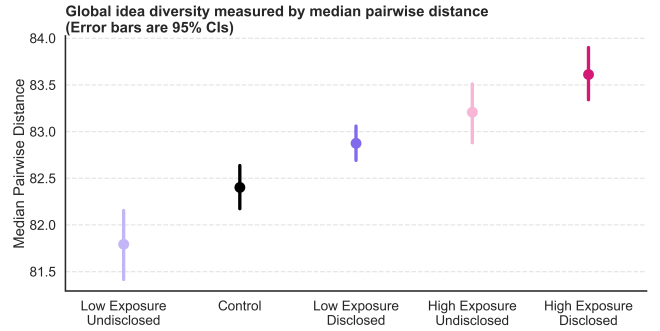
Figure 7: Median pairwise distance of submitted ideas in a condition from Monte Carlo analysis. There was more global diversity of ideas in both high AI exposure conditions than in the control condition.

### 5.2 High AI exposure increases collective diversity

Intuitively, global idea diversity measures how different submitted ideas are in each condition. We measured global idea diversity by looking at the median pairwise distance between neural embeddings of the pool of ideas that were submitted by participants in each condition (Appendix B.2). As a non-parametric measure of effect size, we calculate Cliff's Delta (δ) of these pairwise differences, which ranges from -1 to 1. A value of 0 indicates no difference between the two conditions, +1 indicates values from the first condition are always larger, and -1 indicates the opposite.

We find (Figure 7) that both the [*High Exposure, Disclosed*] ($\delta = 0.31; p = 0.001$) and [*High Exposure, Undisclosed*] ($\delta = 0.26; p = 0.001$) conditions had more global idea diversity than the control

Table 3: Sources of participants and trials. For analysis, we categorized each source into a higher-level interest group (technology, creativity, neutral).

Interest Group	source	Participants (N, % of total)	Trials (N, % of total)
creative	Creative Mornings newsletter	343 (40.6%)	1470 (43.1%)
technology	r/InternetIsBeautiful	298 (35.3%)	1115 (32.7%)
neutral	r/samplesize	94 (11.1%)	389 (11.4%)
neutral	share	61 (7.2%)	250 (7.3%)
technology	r/chatgpt	19 (2.3%)	79 (2.3%)
creative	r/writing	7 (0.8%)	30 (0.9%)
neutral	other	6 (0.7%)	22 (0.6%)
technology	r/singularity	6 (0.7%)	13 (0.4%)
technology	r/artificial	5 (0.6%)	24 (0.7%)
creative	r/poetry	3 (0.4%)	15 (0.4%)
neutral	facebook	2 (0.2%)	7 (0.2%)
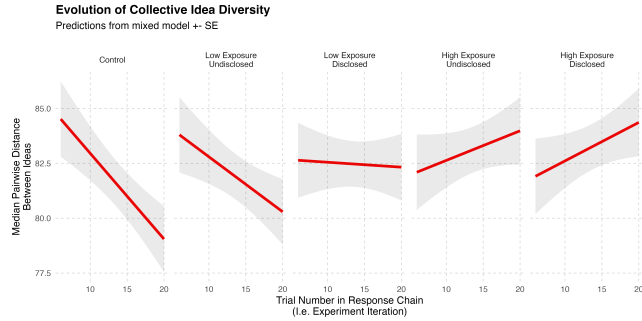
Figure 8: High exposure to AI ideas increased the rate of change in idea diversity. 'Idea diversity' was measured as the cosine distance between submitted ideas.

condition. But of the low exposure conditions, only the [*Low Exposure, Undisclosed*] ($\delta = 0.11$; p = 0.001) condition had higher global diversity than the control condition, with a much smaller effect size than the high exposure conditions. P-values were computed with permutation tests incorporating a conservative adjustment [54] (adding 1 to numerator and denominator) and a Holm-Bonferroni correction for multiple comparisons. As robustness checks, we ran this analysis with alternative measures of idea diversity (SI 8) and the results are directionally similar. We report results for the median pairwise distance as it provides the most *conservative* estimate.

5.3 High AI exposure increases the speed at which idea diversity develops

We measured the evolution of diversity by looking at the diversity (Appendix B.3) of submitted ideas at each iteration in the experiment and seeing if—in some conditions—diversity developed at a faster rate. We find that high AI exposure increases not only the average levels (Figure 7) of collective idea diversity but also the rate of change in idea diversity (Figure 8). As with global idea diversity, different metrics yielded similar regression coefficients (SI 8.3). In the control condition, idea diversity decreased over trials ($\beta = -0.39$, $t(349) = -2.23$, 95% CI = $[-0.73, -0.05]$, $p = 0.03$). That

is, submitted ideas were becoming more similar to each other as the experiment went on. Relative to the control condition, however, the slope of idea diversity with respect to trial number was more positive for the [*High Exposure, Undisclosed*] condition ($\beta = 0.53$, $t(349) = 2.2$, 95% CI = $[0.06, 0.99]$, $p = 0.03$) and the [*High Exposure, Disclosed*] condition ($\beta = 0.57$, $t(349) = 2.37$, 95% CI = $[0.1, 1.03]$, $p = 0.02$). The rate of change in idea diversity for the low AI exposure conditions did not differ from the rate of change in the control condition. Thus, we conclude that high exposure to AI ideas increased the rate of idea diversity relative to the no-AI, control condition.

5.4 Seeing AI ideas does not affect creativity

Conditions did not differ by creativity of ideas, $F(4, 19.86) = 0.12$, $p = 0.97$. We used the *emmeans*[5] joint-tests function to assess main effects of condition after a mixed model (Appendix B.1 for modeling details; SI 9 for regression results). Similarly, no individual treatment condition coefficient differed from zero where the baseline was the control condition. We also tested whether the *evolution* of creativity differed by condition. We conducted a likelihood ratio test on whether adding a [trial number (as an integer) X condition] interaction would improve the fit of a model. The likelihood ratio test indicated that allowing for this temporal interaction *did not* significantly improve the model fit ($\chi^2(4) = 6.52$, $p = 0.16$)—suggesting conditions did not differ on average *or* with respect to the temporal evolution of creativity. All results are based on mixed models (Appendix B.1) with crossed and nested intercepts accounting for non-independence in the experiment.

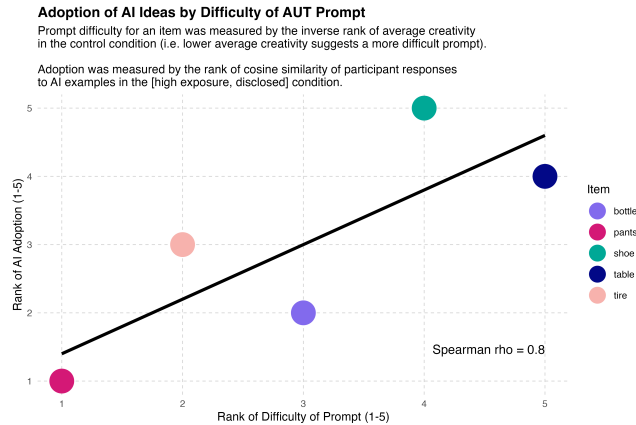## 5.5 How AI Adoption differs depending on task and self-perceived creativity

As a secondary outcome to creativity and diversity, we measured whether AI adoption (the extent to which people submit similar ideas to AI examples) differs depending on the task and person. Prior work found that people may rely more on AI for more difficult tasks [27, 67]. Other work posited that attitudes towards a technology are a major factor when studying human-AI collaboration.

---

[5]https://cran.r-project.org/web/packages/emmeans/index.html

[26]. We measured three attitudinal variables: Attitude toward AI, self-perceived creativity, and perceived creativity of AI. For our individual-level model of adoption, we only included variables as moderators if they significantly improved the fit of a no-moderator, baseline model (Appendix B.1). We operationalized AI adoption as the maximum similarity (following Roemmele [67]) between a submitted idea and AI example ideas a participant saw. The logic for using the maximum is that insofar as a person is influenced by an idea, it would likely be a single one [67].

AI Adoption by Condition and Self-Perceived Creativity
Predictions from estimated marginal means +- SE
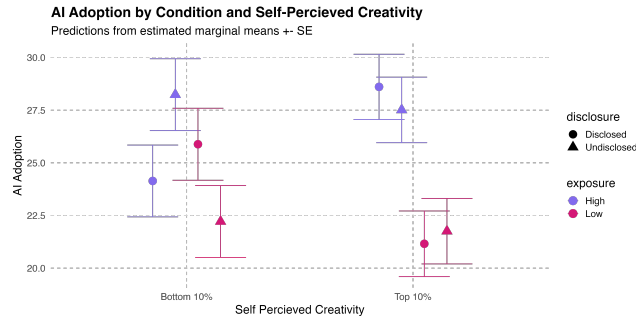
Figure 9: Higher creativity participants adopted AI ideas based on exposure and were not influenced by AI disclosure (i.e., whether AI ideas were labeled as such). 'AI Adoption' was measured by the similarity of a submitted idea to AI example ideas.

Adoption of AI Ideas by Difficulty of AUT Prompt
Prompt difficulty for an item was measured by the inverse rank of average creativity in the control condition (i.e. lower average creativity suggests a more difficult prompt).

Adoption was measured by the rank of cosine similarity of participant responses to AI examples in the [high exposure, disclosed] condition.

Figure 10: When AI ideas were disclosed as such, participants were more likely to adopt AI ideas for more difficult prompts.

5.5.1 People who identify as creative are less influenced by AI disclosure. There was a significant interaction between self-perceived human creativity and the [*High Exposure, Disclosed*] condition ($\beta = 0.11$, $t(2588) = 3.93$, 95% CI = $[0.06, 0.17]$, $p = 0.0001$; SI 10). To probe this interaction, we used our model to predict the estimated marginal mean AI adoption by condition, for both the top 10% and bottom 10% of participants by self-perceived creativity (Figure 9). For high-creativity participants, the effect of exposure

on adoption does not differ by whether AI ideas are disclosed as such: There is no difference in adoption between ([*High Exposure, Undisclosed*] - [*Low Exposure, Undisclosed*]) and ([*High Exposure, Disclosed*] - [*Low Exposure, Disclosed*]), $\Delta = -1.69$, $d = -0.14$, $p = 0.59$. But for low-creativity participants, the difference in adoption for the *undisclosed* conditions ([*High Exposure, Undisclosed*] - [*Low Exposure, Undisclosed*]) was larger than the equivalent difference in adoption for *disclosed* conditions ([*High Exposure, Disclosed*] - [*Low Exposure, Disclosed*]), $\Delta = 7.77$, $d = 0.65$, $p = 0.03$. In short, disclosing ideas as from AI reduced the effect of AI exposure on adoption for lower creativity participants but not higher creativity participants. P-values on post-hoc estimated marginal means were Holm-Bonferroni adjusted for multiple comparisons.

5.5.2 Participants may adopt AI ideas more for harder prompts. We find that when AI ideas are disclosed, participants are more likely to adopt the ideas of AI for difficult AUT prompts (Figure 10). The rank-order correlation between difficulty and adoption was $\rho = 0.8$ for the [*High Exposure, Disclosed*] condition but only $\rho = 0.3$ for the [*High Exposure, Undisclosed*] condition. That is, when people were told ideas were from AI, they were more likely to adopt AI ideas if the prompt was difficult. But since we employed only five items, this finding is speculative. We encourage future work to follow up on this finding.

6 Discussion

Against the backdrop of widespread LLM exposure, we asked: How does exposure to ideas generated by LLMs affect the creativity, diversity, and evolution of human ideas? To answer this, we conducted a large-scale experiment where participants submitted ideas in response to the Alternate Uses Task (a measure of creativity where people brainstorm novel uses of an item) after seeing a set of example ideas. The human examples were from prior participants in the same experimental condition, and the AI ideas were from ChatGPT. The evolving aspect of our experiment, that ideas in a condition feed forward to subsequent trials in that condition, captures the fact that ideas build upon prior ideas. This lets us approximate the cultural evolution effects of AI. Here are three takeaways from our experiment.

6.1 AI makes ideas different but not better

Most notably, exposure to AI ideas did not, on average, make human ideas any 'better' or 'worse' (by creativity). Our high-powered, null finding around average creativity by condition can inform debates about the effect of AI ideas on individual human creativity. Maybe there is little effect. Of course, our experiment is measuring just a single task. But these results suggest that perhaps both worry and optimism around the effect of AI ideas on individual human creativity should be tempered.

Our null finding around creativity contrasts with some prior work suggesting human-AI co-creation enhances the quality of creative outputs [33, 45, 67, 83]. But our study differs from prior studies in its aim and design: We test *passive* exposure to *off-the-shelf* LLMs—not *active engagement* with *optimized-for-creativity* AI aides. The latter is useful for understanding how AI *could* affect creativity. But we aim to approximate how ordinary, existing, and

pervasive AI tools *do* affect the creativity of ideas. At least for this task, we find no evidence of such an effect.

On the other hand, the presence of AI ideas increased the *diversity* of human ideas. This is consistent with work suggesting collaborating with AI leads to more diverse or unexpected outputs [11, 25, 41, 57, 78] and inconsistent with other work that finds collaborating with LLMs decreases diversity [2, 18, 20]. But we highlight that our study is testing *passive exposure* to AI ideas and not *active engagement* with AI ideas. Our setup—passive exposure to AI ideas, scattered amongst human ones—maps onto how many users now experience AI ideas. Hence, it may be that active engagement with LLMs decreases content diversity but simply seeing these ideas as "sparks" [26] *increases* content diversity. And because many more people may be passively exposed to LLM outputs than actively engaging with LLMs, the effect of passive exposure is important to understand.

Another key difference between our study and many similar studies is the brainstorming context. Prior studies typically examine individual participants working directly with AI systems. Our study simulates group brainstorming where participants see ideas from prior group members. In group contexts, AI ideas may serve as additional "voices" that prevent convergence in human-only groups. This suggests AI's effect on ideas may depend on whether it's used for individual ideation versus collective brainstorming. But since ideas we see affect the ideas we create [52], there is an element of collective (in a broad sense) brainstorming even to individual brainstorming.

Crucially, high AI exposure increased both average amounts of diversity *and* rates of change in idea diversity. The latter result is especially important. Small differences in rates of change can accumulate to create large aggregate differences over time. This points to a dynamic that may be missed by static experimental designs. Future work—both simulations and dynamic experiments—can explore the implications of this increase in collective idea diversity unaccompanied by an increase in average individual creativity. For instance, can this dynamic generate "innovation"?

Our finding around the evolution of diversity (Figure 8) is instructive and points to a likely mechanism: Seeing other people's ideas reduced idea diversity in the control condition over time, as successive participants were "converging" on a particular idea sequence. But then injecting AI ideas into the example set increased the diversity of submitted responses by partially "resetting" this convergence. This suggests off-the-shelf LLMs may be viewed as extensive, accessible idea banks that can counter groupthink in collaborative settings. Future research should examine the dynamics and boundary conditions of this potential mechanism's effectiveness.

6.2 High creativity people are less influenced by the source label of ideas

For participants who viewed themselves as highly creative, adoption was driven by exposure and not disclosure (i.e., knowing ideas were from AI). But for lower-creativity participants, knowing the source of an idea *did* affect the adoption of that idea. Perhaps people high in self-reported creativity relied less on source cues when adopting ideas because they were more confident in their ability to judge an idea's creative merit. Regardless, our results suggest that (self-reported) creative people will adopt ideas on the basis of their content. Knowing the source does not matter. In a world where humans have difficulty distinguishing if the content was AI-generated [37], these findings suggest people high in (self-reported) creativity may not be 'duped' into adopting AI ideas.

6.3 Participants may adopt AI ideas when the task is difficult

When AI ideas were labeled, participants were more likely to adopt AI ideas for difficult prompts rather than easy prompts. This finding is similar to what Roemmele [67] observed, where seeing AI examples only influenced creative output when the task was difficult. Both findings are consistent with a theoretical account of task difficulty being associated with increased reliance on automation [27]. However, we emphasize that this finding should be taken as speculative (due to the small number of items) and not conclusive. We encourage future work to test whether users adopt AI ideas for more challenging creative tasks.

If it is the case that users turn to AI for difficult rather than trivial tasks, this would have several implications. While AI can potentially augment humans when they are stuck, researchers raised concerns over 'model collapse' [71]—the deteriorating performance of LLMs when trained on their outputs. If reliance on AI for creative tasks becomes routine, this may contribute to model collapse.

6.4 Limitations And Future Work

Our study has limitations that should be noted and can inform future work. First, we measured the effect of AI ideas for a *single task*. We chose this task because it is one of the most common creativity tasks [1]. But future work could explore other kinds of tasks—especially more complex ones.

Second, we had to operationalize 'AI ideas' in some concrete way (among a large array of choices), and future work can use a similar design but operationalize 'AI ideas' in a different way. We chose a commonly-used LLM at the time of the study, and our prompt was driven by ecological validity and prior work: We used a zero-shot prompt because that is what *users* would likely use, and the specific prompt we used was derived from prior research. We chose not to vary prompts because we didn't want to increase the complexity of an already complex experiment. Future work could explore whether different AI elicitation procedures yield different results. This could involve different models, prompts, LLM-based ideation systems (e.g., [61, 66]), or "simulated social ensembles" [3]. A specific design choice we made is that LLMs were not shown any example stimuli, but humans were. We did this to (A) reflect real-life LLM usage (which is often zero-shot, not few-shot) and (B) conceptually represent LLMs as a 'bank' of ideas that can break up convergence during group brainstorming. However, this specific operationalization leaves open the question of how LLMs versus humans may come up with different ideas if they were given the same example stimuli.

Third, our finding about AI adoption and task difficulty is based on only five AUT items (tire, pants, shoe, table, bottle) and should be considered speculative, not definitive. We encourage future work to explore this hypothesis further. Fourth, future work should test

if alternative classifiers or ways of conceiving variables yield different results. For idea diversity and AI adoption, we addressed this problem by showing that conceptually similar ways of measuring variables yielded qualitatively similar results. For our creativity measure, we used a highly accurate classifier (correlation with human judgments greater than $r = 0.88$ for AUT items we used) trained for this *exact task*, for these *exact AUT items* (Appendix A.1). But of course, all models have some error, and future work based on this model propagates these errors. Incidentally, human judges of creativity only correlate with other human judges at $r = 0.88$ [56], suggesting the classifier we used may be approaching 'the approximate ceiling at which we could expect a model to correlate with human judgments' [56, pg. 11] of creativity. Fifth, the classifier measures one facet of creativity: originality. Future work can also explore whether AI ideas have different effects on other facets of creativity such as fluency (how many ideas) or elaboration (how detailed an idea is). Sixth, we employed a non-representative sample of technology-interested users and creative professionals. While these two groups were most relevant to the phenomenon in question, our sample also limits generalizability. Seventh, our design captures some aspects of human-AI cultural evolution but not others; it is a stylized model that does not capture all the complexity of how AI can shape collective behavior [12, 23]. For example, future work could explore more complex transmission scenarios.

Despite these limitations, our work offers a large-scale, dynamic account of how seeing LLM ideas affects collective thought. This work opens the door to more dynamic experiments exploring the compounding effects of "AI in the culture loop".

6.5 Conclusion: Passive exposure to AI ideas affects collective thought

Our experiment suggests that passive exposure to AI ideas—the kind of passive exposure we are now *inundated with* in a post-ChatGPT era—does affect collective thought. Even small effects are meaningful. This exposure is both pervasive and growing. But the effects of AI ideas are nuanced. Seeing AI ideas did not increase individual creativity, though it did increase collective diversity. The effects of AI ideas vary across individuals and tasks. There is still much to learn. We hope our study inspires more research on how passive exposure to AI ideas affects collective thought via the use of multiple-worlds experiments.

Acknowledgments

References

[1] Anna Abraham. 2016. Gender and creativity: an overview of psychological and neuroscientific literature. *Brain Imaging and Behavior* 10, 2 (June 2016), 609–618. doi:10.1007/s11682-015-9410-8

[2] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Creativity and Cognition*. ACM, Chicago IL USA, 413–425. doi:10.1145/3635636.3656204

[3] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A System for Guiding LLMs via Simulated Social Ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3706598.3713675

[4] Raiyan Abdul Baten, Richard N. Aslin, Gourab Ghoshal, and Ehsan Hoque. 2021. Cues to gender and racial identity reduce creativity in diverse social networks. *Scientific Reports* 11, 1 (May 2021), 10261. doi:10.1038/s41598-021-89498-5

[5] Roger E. Beaty and Dan R. Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods* 53, 2 (April 2021), 757–780. doi:10.3758/s13428-020-01453-w

[6] Roger E. Beaty, Dan R. Johnson, Daniel C. Zeitlen, and Boris Forthmann. 2022. Semantic Distance and the Alternate Uses Task: Recommendations for Reliable Automated Assessment of Originality. *Creativity Research Journal* 34, 3 (July 2022), 245–260. doi:10.1080/10400419.2022.2025720

[7] Roger E. Beaty, Daniel C. Zeitlen, Brendan S. Baker, and Yoed N. Kenett. 2021. Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. *Thinking Skills and Creativity* 41 (September 2021), 100859. doi:10.1016/j.tsc.2021.100859

[8] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[9] Charles F. Bond and Linda J. Titus. 1983. Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin* 94, 2 (1983), 265–292. doi:10.1037/0033-2909.94.2.265

[10] Robert Boyd and Peter J. Richerson. 1988. *Culture and the Evolutionary Process*. University of Chicago Press.

[11] Boyd Branch, Piotr Mirowski, and Kory W. Mathewson. 2021. Collaborative Storytelling with Human Actors and AI Narrators. http://arxiv.org/abs/2109.14728

[12] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L. Griffiths, Joseph Henrich, Joel Z. Leibo, Richard McElreath, Pierre-Yves Oudeyer, Jonathan Stray, and Iyad Rahwan. 2023. Machine culture. *Nature Human Behaviour* 7, 11 (Nov. 2023), 1855–1868. doi:10.1038/s41562-023-01742-2 Publisher: Nature Publishing Group.

[13] Vincent R. Brown and Paul B. Paulus. 2002. Making Group Brainstorming More Effective: Recommendations From an Associative Memory Perspective. *Current Directions in Psychological Science* 11, 6 (December 2002), 208–212. doi:10.1111/1467-8721.00202

[14] Jason W. Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A. Bakker, Joshua A. Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, Lucie Flek, Stefan M. Herzog, Saffron Huang, Sayash Kapoor, Arvind Narayanan, Anne-Marie Nussberger, Taha Yasseri, Pietro Nickl, Abdullah Almaatouq, Ulrike Hahn, Ralf H. J. M. Kurvers, Susan Leavy, Iyad Rahwan, Divya Siddarth, Alice Siu, Anita W. Woolley, Dirk U. Wulff, and Ralph Hertwig. 2024. How large language models can reshape collective intelligence. *Nature Human Behaviour* 8, 9 (September 2024), 1643–1655. doi:10.1038/s41562-024-01959-9

[15] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.

[16] Ted Chiang. 2023. ChatGPT Is a Blurry JPEG of the Web. *The New Yorker* (February 2023). https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web

[17] Hao Cui and Taha Yasseri. 2024. AI-enhanced collective intelligence. *Patterns* 5, 11 (November 2024). doi:10.1016/j.patter.2024.101074

[18] Fabrizio Dell'Acqua, Edward McFowland III, 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. doi:10.2139/ssrn.4573321

[19] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition (C&C '22)*. Association for Computing Machinery, New York, NY, USA, 623–627. doi:10.1145/3527927.3535197

[20] Anil R. Doshi and Oliver P. Hauser. 2024. Generative artificial intelligence enhances creativity but reduces the diversity of novel content. doi:10.48550/arXiv.2312.00506

[21] Denis Dumas, Peter Organisciak, Shannon Maio, and Michael Doherty. 2021. Four Text-Mining Methods for Measuring Elaboration. *The Journal of Creative Behavior* 55, 2 (2021), 517–531. doi:10.1002/jocb.471

[22] Tojin T. Eapen, Daniel J. Finkenstadt, Josh Folk, and Lokesh Venkataswamy. 2023. How Generative AI Can Augment Human Creativity. *Harvard Business Review* (July 2023). https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity

[23] Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. Large AI models are cultural and social technologies. *Science* 387, 6739 (March 2025),

1153–1156. doi:10.1126/science.adt9819 Publisher: American Association for the Advancement of Science.

[24] Katy Ilonka Gero. 2023. *AI and the Writer: How Language Models Support Creative Writers*. Ph.D. Columbia University, United States – New York. https://www.proquest.com/docview/2753687892/abstract/ACF7F21F1E274995PQ/1

[25] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. doi:10.1145/3290605.3300526

[26] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference (DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1002–1019. doi:10.1145/3532106.3533533

[27] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2014. Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics* 83, 5 (May 2014), 368–375. doi:10.1016/j.ijmedinf.2014.01.001

[28] Andrew Griffin. 2024. ChatGPT creators OpenAI are generating 100 billion words per day, CEO says. *The Independent* (February 2024). https://www.independent.co.uk/tech/chatgpt-openai-words-sam-altman-b2494900.html

[29] J.P. Guilford. 1967. *The nature of human intelligence*. McGraw-Hill, New York, NY, US.

[30] Joy Paul Guilford. 1978. *Alternate uses*. Sheridan supply Company.

[31] Erik E. Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: AI takes the Torrance Test. *Journal of Creativity* 33, 3 (December 2023), 100065. doi:10.1016/j.yjoc.2023.100065

[32] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 89–100. doi:10.1093/jcmc/zmz022

[33] Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. 2022. Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior* (October 2022), 107502. doi:10.1016/j.chb.2022.107502

[34] Krystal Hu. 2023. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters* (February 2023). https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[35] Chieh-Yang Huang, Shih-Hong Huang, and Ting-Hao Kenneth Huang. 2020. Heteroglossia: In-Situ Story Ideation with the Crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376715

[36] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. doi:10.1145/3411764.3445270

[37] Maurice Jakesch, Jeffrey Hancock, and Mor Naaman. 2022. Human Heuristics for AI-Generated Language Are Flawed. doi:10.48550/arXiv.2206.07271

[38] Jared Henderson. 2022. ChatGPT Will Make You Less Creative. https://www.youtube.com/watch?v=1K8PiMNoR7A

[39] Ulrich Kirk, Martin Skov, Oliver Hulme, Mark S. Christensen, and Semir Zeki. 2009. Modulation of aesthetic value by semantic context: An fMRI study. *NeuroImage* 44, 3 (February 2009), 1125–1132. doi:10.1016/j.neuroimage.2008.10.009

[40] Krish Naik. 2023. Will Chatgpt Kill Your Creativity? https://www.youtube.com/watch?v=0m2r9elReBY

[41] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. doi:10.1145/3491102.3502030

[42] Jasmine Mangalaseril. 2023. The Incredible Blandness of ChatGPT. https://cardamomaddict.substack.com/p/the-incredible-blandness-of-chatgpt

[43] Arthur I. Miller. 2019. *The Artist in the Machine: The World of AI-Powered Creativity*. The MIT Press. https://direct.mit.edu/books/book/4547/The-Artist-in-the-MachineThe-World-of-AI-Powered

[44] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–34. doi:10.1145/3544548.3581225

[45] Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4918–4929. doi:10.18653/v1/2020.findings-emnlp.442

[46] Kathleen L. Mosier, Linda J. Skitka, Mark D. Burdick, and Susan T. Heers. 1996. Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, 4 (October 1996), 204–208. doi:10.1177/154193129604000413

[47] Michael D. Mumford and Sven Hemlin. 2017. *Handbook of Research on Leadership and Creativity*. Edward Elgar Publishing.

[48] Reem Nadeem. 2022. How Americans think about artificial intelligence. https://www.pewresearch.org/internet/2022/03/17/how-americans-think-about-artificial-intelligence/

[49] Reem Nadeem. 2023. Public Awareness of Artificial Intelligence in Everyday Activities. https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/

[50] Nation World News. 2023. Why Does ChatGPT Increase Creativity? https://nationworldnews.com/why-does-chatgpt-increase-creativity/

[51] J. Nickerson and Yasuaki Sakamoto. 2010. Crowdsourcing Creativity: Combining Ideas in Networks. https://www.semanticscholar.org/paper/Crowdsourcing-Creativity%3A-Combining-Ideas-in-Nickerson-Sakamoto/340a7645d1402287e151e83981f8a4085227e317

[52] Bernard A. Nijstad and Wolfgang Stroebe. 2006. How the Group Affects the Mind: A Cognitive Model of Idea Generation in Groups. *Personality and Social Psychology Review* 10, 3 (August 2006), 186–213. doi:10.1207/s15327957pspr1003_1

[53] American Psychological Association Dictionary of Psychology. [n. d.]. divergent thinking. https://dictionary.apa.org/divergent-thinking

[54] Markus Ojala and Gemma C. Garriga. 2009. Permutation Tests for Studying Classifier Performance. In *2009 Ninth IEEE International Conference on Data Mining*. IEEE, Miami Beach, FL, USA, 908–913. doi:10.1109/ICDM.2009.108

[55] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2022. Beyond Semantic Distance: Automated Scoring of Divergent Thinking Greatly Improves with Large Language Models. *Researchgate (v1)* (August 2022). doi:10.13140/RG.2.2.32393.31840

[56] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49 (September 2023), 101356. doi:10.1016/j.tsc.2023.101356

[57] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–10. doi:10.1145/3411763.3450391

[58] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity? doi:10.48550/arXiv.2309.05196

[59] Paul B. Paulus and Vincent R. Brown. 2007. Toward More Creative and Innovative Group Idea Generation: A Cognitive-Social-Motivational Perspective of Brainstorming. *Social and Personality Psychology Compass* 1, 1 (2007), 248–265. doi:10.1111/j.1751-9004.2007.00006.x

[60] Marieke M. M. Peeters, Jurriaan van Diggelen, Karel van den Bosch, Adelbert Bronkhorst, Mark A. Neerincx, Jan Maarten Schraagen, and Stephan Raaijmakers. 2021. Hybrid collective intelligence in a human–AI society. *AI & SOCIETY* 36, 1 (March 2021), 217–238. doi:10.1007/s00146-020-01005-y

[61] Kevin Pu, K. J. Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2025. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–31. doi:10.1145/3706598.3714057

[62] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. http://arxiv.org/abs/1908.10084

[63] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1364–1378. doi:10.1145/2675133.2675246

[64] European Business Review. 2023. ChatGPT: Ushering in the Age of Creativity. https://www.europeanbusinessreview.com/chatgpt-ushering-in-the-age-of-creativity/

[65] Peter J. Richerson and Robert Boyd. 2008. *Not By Genes Alone: How Culture Transformed Human Evolution*. University of Chicago Press.

[66] Steven R. Rick, Gianni Giacomelli, Haoran Wen, Robert J. Laubacher, Nancy Taubenslag, Jennifer L. Heyman, Max Sina Knicker, Younes Jeddi, Hendrik Maier, Stephen Dwyer, Pranav Ragupathy, and Thomas W. Malone. 2023. Supermind Ideator: Exploring generative AI to support creative problem-solving. doi:10.48550/arXiv.2311.01937 arXiv:2311.01937 [cs].

[67] Melissa Roemmele. 2021. Inspiration through Observation: Demonstrating the Influence of Automatically Generated Text on Creative Writing. doi:10.48550/arXiv.2107.04007

[68] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science (New York, N.Y.)* 311, 5762 (2006), 854–856. doi:10.1126/science.1121066

[69] Matthew J. Salganik and Duncan J. Watts. 2009. Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science* 1, 3 (2009), 439–468. doi:10.1111/j.1756-8765.2009.01030.x

[70] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the Influence of Explainable AI on Automation Bias. doi:10.48550/arXiv.2204.08859

[71] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes

Models Forget. http://arxiv.org/abs/2305.17493

[72] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (February 2015), 937–945. doi:10.1145/2675133.2675239

[73] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting GPT-3's Creativity to the (Alternative Uses) Test. doi:10.48550/arXiv.2206.08932

[74] Tubefilter. 2023. 86% of creators believe AI has a positive effect on creativity. ChatGPT offered its own opinions. https://www.tubefilter.com/2023/06/02/lightricks-creator-artificial-intelligence-ai-survey-chat-gpt-wired/

[75] Chetan Walia. 2019. A Dynamic Definition of Creativity. *Creativity Research Journal* 31, 3 (July 2019), 237–247. doi:10.1080/10400419.2019.1641787

[76] Wilcot. 2023. Using Chat-GPT for Innovators: Enhancing Creativity and Innovation. https://www.boardofinnovation.com/blog/using-chat-gpt-for-innovators-enhancing-creativity-and-innovation/

[77] Jamie Williams. 2018. Should AI Always Identify Itself? It's More Complicated Than You Might Think. https://www.eff.org/deeplinks/2018/05/should-ai-always-identify-itself-its-more-complicated-you-might-think

[78] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and L. C. Ray. 2022. AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing 56-65. https://www.semanticscholar.org/paper/AI-as-an-Active-Writer%3A-Interaction-Strategies-with-Yang-Zhou/15ddeb7765e2a3ea692a27d9b30e8f9446d74742

[79] Tianchen Yang, Qifan Zhang, Zhaoyang Sun, and Yubo Hou. 2023. Automatic Assessment of Divergent Thinking in Chinese Language with TransDis: A Transformer-Based Language Model Approach. doi:10.48550/arXiv.2306.14790

[80] Lixiu Yu and Jeffrey V. Nickerson. 2011. Cooks or cobblers? crowd creativity through combination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1393–1402. doi:10.1145/1978942.1979147

[81] Lixiu Yu and Jeffrey V. Nickerson. 2013. An internet-scale idea generation system. *ACM Transactions on Interactive Intelligent Systems* 3, 1 (April 2013), 2:1–2:24. doi:10.1145/2448116.2448118

[82] Yuhua Yu, Roger E. Beaty, Boris Forthmann, Mark Beeman, John Henry Cruz, and Dan Johnson. 2023. A MAD method to assess idea novelty: Improving validity of automatic scoring using maximum associative distance (MAD). *Psychology of Aesthetics, Creativity, and the Arts* (2023), No Pagination Specified–No Pagination Specified. doi:10.1037/aca0000573

[83] 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 (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. doi:10.1145/3490099.3511105

A Stimuli Construction

A.1 Choosing AUT Items

We had to choose a selection of items that people would brainstorm creative uses for. We chose five items for which the creativity classifier that we used had the highest accuracy. Previously, Organisciak et al. [55] fine-tuned GPT-3 Davinci to predict the creativity of AUT items. We use a dataset from Organisciak et al. [55] containing 20,121 responses from 2,025 participants, across 21 distinct AUT items and nine distinct studies [55].[6] Each response was graded for creativity by humans and normalized to a scale of 1-5. Then [55] fine-tuned GPT-3 Davinci to predict creativity scores. Here, fine-tuning involves providing {*Input* (an AUT response), *Output* (human rating)} pairs to a pre-trained LLM. Then the LLM adjusts its parameters to produce a similar output given an input, proxying human judgements. Overall, the fine-tuned GPT-3 classifier had a correlation[7] of $r = 0.81$ with human judgment. Accuracy varied by item (SI 1). For our experiment, we picked the five items for

---

[6]We obtained an early version of this dataset by direct correspondence with Dr. Organisciak on February 23, 2023; the code that Dr. Organisciak used to generate this dataset is available at https://github.com/massivetexts/llm_aut_study/blob/main/notebooks/Process_AUT_GT.ipynb

[7]We obtained scores for this classifier by downloading the zip file from (https://github.com/massivetexts/llm_aut_study/blob/main/results/evaluation.zip), then navigating to `gt_main2/gpt-ft-davinci-1.csv`

which the classifier had the highest accuracy: *tire* (r=0.91), *pants* (r=0.91), *shoe* (r=0.91), *table* (r=0.9), and *bottle* (r=0.88). Our task is "in-domain" for the classifier since we ask participants to do the same task for the same items the classifier was trained on.

A.2 Generating GPT Ideas

Prompt Construction. Our specific zero-shot prompt was informed by prior work on LLMs and creativity. [73] administered the AUT to GPT-3 through a zero-shot prompt. However, this prompt generated much wordier responses ($M = 24.1, SD = 8.8$) than the human responses in the Organisciak Dataset ($M = 4.6, SD = 5$). Such a discrepancy would alert participants to what was AI vs. human generated, which would nullify the disclosure factor (whether the source of an idea is disclosed). Hence, we appended a request (Figure 11) to use roughly the same number of words (5) as the average human response.

Word Length Prompt Experiment. We conducted a Monte Carlo experiment to confirm our modified zero-shot prompt resulted in responses with a similar word length to humans. Using parameters from Stevenson et al. [73]'s experiments: For $n = 1000$ trials, we fixed presence penalty and frequency penalty at 1, randomly chose a temperature (higher values lead to more randomness) in [0.65, 0.7, 0.75, 0.8], and randomly chose one of our 5 AUT items. For each trial, ChatGPT generated five ideas. The modified prompt resulted in responses with an average word length (M=4.44, SD = 1.34) much closer to human responses (M=4.56, SD = 4.97) than the original zero-shot prompt (M = 25.38, SD = 8.55). A permutation test (more details below) further shows that this difference was significant at $p < 0.001$. We used the ideas generated by our modified zero-shot prompt as stimuli for the main experiment.

We conducted a permutation test of whether the limited-length zero-shot prompt ('ZeroShotLimit') yielded more similar word counts to the human responses than the original zero-shot prompt ('ZeroShot') from [73]. Note that permutation tests are non-parametric, so they are robust to (e.g.) violations of normality. The observed difference is computed as:

$$T^{obs} =$$

$$|Avg(\text{ZeroShot}) - Avg(\text{HumanResponses})| -$$

$$|Avg(\text{ZeroShotLimit}) - Avg(\text{HumanResponses})|$$

where Avg is the average word count of a response and the human responses are from [55].

Under the null hypothesis, we assume prompts do not differ in respective word counts, so we can swap the labels of 'ZeroShot' and 'ZeroShotLimit'. To simulate draws from this null distribution, we permuted the labels of the source column 10000 times, each time calculating a new value for the difference of differences defined in T. The p-value is the proportion of permutations where the null T is greater than or equal to the observed difference, T^{obs}. As in [54], we apply a conservative adjustment and add 1 to the numerator and denominator (this means we never get a p-value of 0).

$$p = \frac{[\sum_{i=1}^{n} I(T^{null} >= T^{obs})] + 1}{n + 1}$$

```
What are some creative
uses for a [OBJECT]? The
goal is to come up with
creative ideas, which
are ideas that strike
people as clever, unusual,
interesting, uncommon,
humorous, innovative, or
different. List creative
uses for a [OBJECT]. Make
sure each response is [MEAN
HUMAN WORDS] words.
```

Figure 11: To generate AUT ideas, we used the zero-shot prompt from [73] with an additional instruction at the end to match the mean length of human responses from prior work.

This p-value indicates the likelihood of observing a difference as extreme as our observed difference if the null hypothesis—that prompts did not differ in word counts—is true. We conclude the observed difference is significant.

B Outcome Measures

B.1 Local Level

Outcomes at the local level—the level of an individual trial—are useful for two reasons. First, this level shows how a participant's response relates to the examples they see. Second, this level lets us model whether individual differences moderate the effect of either disclosure or exposure. For each of our local outcomes, we have a baseline model that uses crossed random intercepts to account for the multilevel structure of the experiment. The first random intercept is for participants, accounting for clustering due to repeated measures. This random intercept is then crossed with a second random intercept for response chains, which we nest inside of {[*condition*], *item*} combinations. In R syntax, the random effect structure was `... + (1|ParticipantID) + (1|ItemCondition/ResponseChainID)`. Models were fit in the *lme4* R package. We computed profile likelihood confidence intervals for coefficients using the *confint* R package. We used estimated marginal means (*emmeans* R package) to conduct model-adjusted F-tests, linear contrasts, predictions, and pairwise comparisons. We apply Holm-Bonferroni adjustments to pairwise comparison p-values. Our baseline 'local' model is:

$$\text{variable}_{ijk} = \beta_0 + \beta_1\text{condition}_j + \beta_2\text{CreativityHuman}_i +$$
$$\beta_3\text{AiRelCreate}_i + \beta_4\text{AiFeeling}_i +$$
$$\beta_5\text{InterestGroup}_i + \beta_6\text{ConditionOrder}_{ijk} +$$
$$\beta_7\text{LogDuration}_{ijk} + \beta_8\text{nSeedsPresent}_{ijk} +$$
$$\beta_9\text{TrialNo}_{jk} + u_{0i} + v_{0jk} + e_{ijk} \quad (1)$$

where

- i indexes participants, j indexes item-condition combinations, k indexes response chains.

- **CreativityHuman** is self-perceived creativity relative to AI.
- **AiRelCreate** is constructed as (self-perceived creativity to humans) - (self-perceived creativity to AI). Note that this is an implicit measure of AI's (perceived) creativity relative to humans. For example, if you say you are more creative than 40% of humans and 60% of AI, then AiRelCreate = -20, as the implicit belief is that AI is *less creative* (-20 percentile points) than humans. Conversely, if you say you are more creative than 50% of humans but 30% of AI then the implicit belief is that AI is *more creative* (50% - 30% = +20) than humans.
- **AiFeeling** refers to the AI sentiment question.
- **InterestGroup** maps each source of the experiment to categories: creative, neutral, or technology. These categories are described in Table 3.
- **ConditionOrder** denotes the sequence in which the participant was assigned to complete the trial (e.g., the 1st time a participant took part, etc.).
- **LogDuration** is the natural logarithm of the time (in seconds) a participant spent before submitting their answer.
- **nSeedsPresent** controls for the number of examples the participant saw that were seed ideas from the Organisciak Dataset.
- **TrialNo** indicates the trial number within a specific response chain. For example: the 18th response for {[*Control*], *tire*, response chain 5}

We balanced interest in testing experimental hypotheses that conditions differed by subgroups with caution around model overfitting. We considered interactions between the treatment condition and four potential moderators: self-perceived human creativity (*CreativityHuman*), AI - Human creativity (*AiRelCreate*), feeling towards AI (*AiFeeling*), and interest group (*InterestGroup*). We first conducted likelihood ratio tests to test if adding each moderator improved our baseline model. Moderators were kept only if they significantly improved the fit ($p < 0.05$). See SI 7 for retained moderators. Then, we used emmeans to probe and interpret moderating effects.

B.2 Global

Intuitively, the *global* diversity of ideas in a condition measures how similar or different submitted ideas in a condition tend to be. The relevant level of aggregation here is *all* of the submitted ideas at a {[*condition*], *item*} level. For example, consider the total set of ideas participants submitted for a *tire* in the [*Control*] condition. Is this set of ideas more diverse from each other than the set of submitted ideas for a tire in the [*High Exposure*, *Disclosed*] condition?

We used a Monte Carlo procedure and permutation tests to assess if conditions differed with respect to these metrics. For 50 Monte Carlo runs, for each {[*condition*], *item*} combination, we randomly sampled 50 ideas and computed idea diversity metrics. We then conducted pairwise paired (at the level of Monte Carlo seeds and items) permutation tests with 10,000 iterations to see if the two conditions differed on these metrics. As a non-parametric measure of effect size, we also calculate Cliff's Delta (δ), which ranges from -1 to 1. A value of 0 indicates no difference between the two conditions, +1 indicates values from the first condition are

always *larger*, and -1 indicates the opposite. See SI 8.1 for more details.

B.3 Evolution

Creativity. To test if conditions differed in their evolution of creativity, we conducted a likelihood ratio test on whether an interaction between *condition* and *TrialNo* significantly improved the fit of the local creativity model.

Idea Diversity. Intuitively, we are interested in if—as the experiment goes on—ideas that participants submit tend to become more or less similar to each other. We use the trial number in a response chain to index time in the experiment. For example, is the set of submitted responses at trial number 4 more or less similar to each other as the set of submitted responses at trial number 20? Here, the diversity of interest is not between a submitted response and example responses but between all submitted responses at a given 'time point' (i.e., trial number). The question is if the diversity increases or decreases as the experiment goes on and if this rate of change differs by condition. Here are the mechanics of our process. See SI Section 8.3 for more details.

(1) We first 'pooled' together all ideas at the {[*condition*], *item*, *trial number*} level, across response chains. For example, consider all ideas for a *tire* for the [*Control*] condition that were the *fourth* response in a response chain. We refer to this set as a 'pool' of ideas.
(2) We next computed idea diversity measures for each pool of ideas, where idea pools were defined in (1). We use the same metrics that we measure at a local level for idea diversity. Median pairwise distance is our main measure. We conduct robustness checks using mean pairwise distance and mean distance from the centroid. Each metric shows qualitatively similar results.
(3) We then fit a mixed model (items as random intercepts) to test if the slope of trial number on idea diversity differed by condition. That is: Are submitted responses in some conditions changing at a faster rate?

C Effects of Exposure and Disclosure on Creativity and Diversity

C.1 Factor 1: LLM Exposure

C.1.1 Effects on Creativity. Intuitively, the effect of exposure to ChatGPT ideas will depend on how creative ChatGPT answers are relative to human ideas. In preliminary testing, we found that the answers to the AUT generated by our prompt were scored as more creative than the ideas generated by humans (see SI 4) via the Organisciak et al. [56] classifier. LLM generations may be increasing in creativity: while GPT-3 (an earlier model than ChatGPT-3.5) scored lower in AUT creativity than humans on the AUT [73], GPT-4 (a more recent model than ChatGPT-3.5) scored among the *top percentile* of humans on a similar verbal creativity task [31], as measured by human judges.

Even if language models can generate creative ideas, it is unclear from prior work if mere exposure to these ideas can increase human creativity. On one hand, the associative model of brainstorming suggests that exposure to others' ideas can stimulate idea generation

by activating a non-accessible concept of a participant's memory [13, 52, 59]. For example, ChatGPT may come up with a use for a bottle that you never associated with bottles. This can then inspire you to come up with creative uses along this line. In this way, ChatGPT can *stimulate* creativity. On the other hand, there is also evidence that seeing the ideas of others *inhibits* a participant's idea generation if "one is exposed to an idea that has few connections to other ideas in an individual's semantic network" [59, pg. 10]. Indeed, this appeared to be the case in Yang et al. [78]. There is a possibility that AI ideas are creative but so divorced from how humans generate ideas that seeing these ideas actually has an *inhibiting* effect. Separate from prior *academic* work, there are *public* debates about the impact of LLMs (such as ChatGPT) on creativity (e.g., [22, 38, 40, 50, 64, 74, 76]). Many of these debates *assume* ChatGPT will have *some* impact on an individual's creativity—either good or bad. Our work contributes empirical results to this broader public conversation.

C.1.2 Effects on Diversity. Prior work in AI co-creation finds mixed effects. Collaborating with AI can lead to more diverse [11, 25, 41, 57, 78] or less diverse outputs [18, 20, 58]. But note that these studies are testing *active engagement*, and most test active engagement with *intentionally* constructed systems. This is different from the passive, incidental exposure to AI ideas that now occur in everyday life. Writers call ChatGPT 'a blurry JPEG of the internet' [16] and discuss its 'incredible blandness' [42]; researchers call it a 'stochastic parrot' [8]. It is not clear, then, how passive exposure to ideas from off-the-shelf LLMs—precisely the kind we are inundated with—would affect the diversity of human ideas.

C.2 Factor 2: LLM Disclosure

C.2.1 Effects on Creativity. Building on Hwang and Won [36], we employ the theory of social facilitation [9] to understand how LLM disclosure can affect human creativity. Facilitation theory is concerned with how the presence of others affects one's performance. Hwang and Won [36] asked participants to brainstorm with chatbots (which gave pre-programmed responses) and experimentally varied whether or not participants were told that their partner was a chatbot. Disclosing that the partner was a chatbot led to higher creativity in participant responses, which Hwang and Won [36] attributes to the novelty of brainstorming with a chatbot. We build on this notion of *facilitation* as a theoretical lens. However, it is not clear if Hwang and Won [36]'s finding (that telling people they are brainstorming with a chatbot increases creativity) would replicate in our study, especially in a post-ChatGPT era. First, we are measuring *exposure* and not direct engagement with chatbots. The novelty of a chatbot may be higher when you are the one working with it to generate ideas. Second, presumably, the novelty of talking to a chatbot may be lower due to the widespread popularity of ChatGPT. Moreover, we may expect heterogeneity in disclosure's effect on creativity and diversity. It may be that users who have lower self-perceived creative abilities may feel 'competition' with AI due to its presence and, in turn, submit more creative responses when they know the ideas they are exposed to are from AI.

C.2.2 Effects on Diversity. It is not clear how knowing content is from AI will affect the diversity of ideas participants produce.

But prior work suggests heterogeneity along two lines: the difficulty of the prompt[8], and the attitude of the participant. Prior work suggests that disclosing ideas as AI-generated would *decrease* diversity due to automation bias, the tendency to over-rely on AI systems [27, 46, 70]. Increased reliance on AI ideas (when labeled as such) could lead to lower idea diversity and higher AI adoption. Conversely, some evidence suggests people display algorithmic aversion to creative products such as haikus [33] or art [39]. This aversion would yield the opposite prediction. Roemmele [67] found that seeing AI examples only affected the participant's writing on a key measure for difficult prompts—suggesting creative task difficulty might moderate the effect of disclosure on AI adoption. Task confidence decreases reliance on automated systems and trust in a system increases reliance on automated systems [27]. Although this literature is not usually applied to creativity, we might then suspect that self-reported creativity (i.e., roughly analogous to task confidence) and perceived AI creativity could affect the adoption of AI ideas when the source is disclosed.

D Experiment Screenshots

Figure 12: Screenshot of a trial participants are told to complete. This is the [*Low Exposure, Disclosed*] condition since there are two AI ideas and the AI ideas are labeled.

Figure 13: After each trial, participants were given feedback as motivation to continue.

---

[8]We measure the 'difficulty' of a prompt by the inverse rank of the average creativity in the control condition. If participants tended to submit *lower* creativity ideas in the control condition for item X, we said item X was *difficult*.