Exploring the Impact of AI Value Alignment in Collaborative Ideation: Effects on Perception, Ownership, and Output

by

Alicia Guo

S.B. in Computer Science and Engineering
Massachusetts Institute of Technology (2022)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by:	Alicia Guo Department of Electrical Engineering and Computer Science September 12, 2023
Certified by:	Pattie Maes Professor of Media Arts and Sciences Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Exploring the Impact of AI Value Alignment in Collaborative Ideation: Effects on Perception, Ownership, and Output

by

Alicia Guo

Submitted to the Department of Electrical Engineering and Computer Science
on September 12, 2023, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

AI-based virtual assistants are increasingly used to support daily ideation tasks. The values or bias present in these agents can influence output in hidden ways. They may also affect how people perceive the ideas produced with AI agents of different value alignments and lead to implications for the design of AI-based tools. We explored the effects of AI agents with different values on the ideation process and user perception of idea quality, ownership, agent competence, and values present in the output. Our study tasked 180 participants with brainstorming practical solutions to a set of problems with AI agents of different values. Results show no significant difference in self-evaluation based on value alignment; however, the ideas generated in the brainstormig process reflected the AI's values. This thesis highlights an intricate interplay between AI values and human ideation, suggesting careful design considerations for future AI-supported brainstorming tools.

Thesis Supervisor: Pattie Maes
Title: Professor of Media Arts and Sciences

Acknowledgments

I would like to thank my thesis mentor and coach, Pat Pataranutaporn, who was an incredible mentor and supported my thesis and ideas every step of the way. Your enthusiasm and research guidance have shaped me as a researcher. I would also like to thank my thesis advisor, Professor Pattie Maes for her invaluable insights and express my gratitude to the Fluid Interfaces group for welcoming me this year. Finally, I am most grateful to my friends and family, and especially my mom and dad for always supporting me.

Contents

1 Introduction **13**

2 Related Work **15**

2.1 Ideation process 15

2.2 AI assistants and tools for creativity 16

2.3 Opinionated AI 17

2.4 Interactions with AI for agency and ownership 18

3 Methodology **21**

3.1 Study overview 21

3.2 Research questions 23

3.3 Value domains 23

3.4 AI agent interaction 26

3.5 Measurements 27

3.6 Analysis 29

4 Results **31**

4.1 Values in brainstorming output 31

4.2 Self-evaluation questions 33

4.3 Brainstorming feedback 35

5 Discussion **39**

5.1 Implications 39

5.2 Limitations 40

5.3 Future work 40

6 Conclusion **41**

A Brainstorming task example **43**

List of Figures

1-1 Study summary: examining the effects of user value alignment with opinionated AI agents on brainstorming outputs 13

3-1 Study flow 22
3-2 Distribution of opinions on potential domains 25
3-3 AI brainstorming agent 26

4-1 Output values split by AI value and participant values 31
4-2 Output values further split by domain 33
4-3 Significant post-survey questions across human-AI conditions 34

List of Tables

3.1 Values for each domain 24

Chapter 1

Introduction

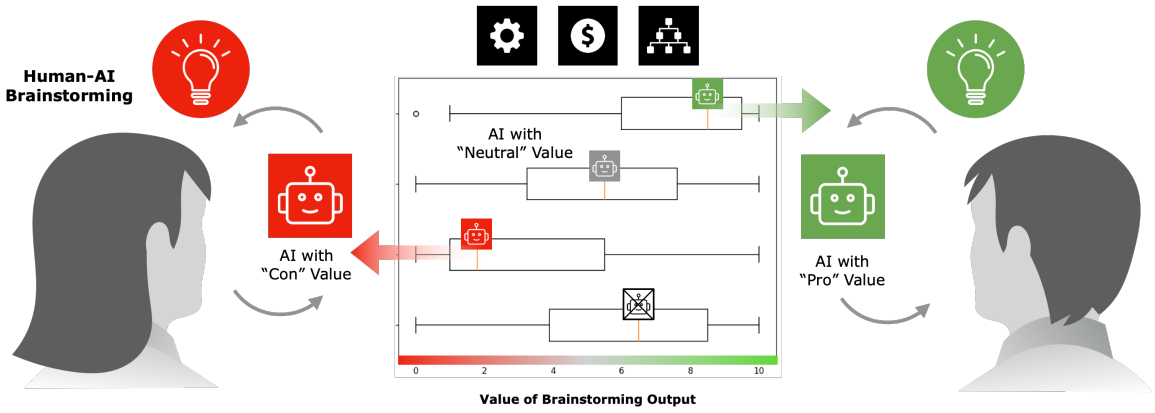

Figure 1-1: Study summary: examining the effects of user value alignment with opinionated AI agents on brainstorming outputs

In recent years, using large language models in everyday tasks such as assistants for writing, programming, and customer service has become increasingly common. With the release of GPT-4, AI assistants have been further integrated into the workflows of common tasks. This technology has allowed for increased productivity and convenience, however, the long-term effects of these new work patterns on a person and the output are still unknown. Much of the focus has been on the efficiency and outputs of using AI systems, without enough emphasis on the process itself and what that means for the user's sense of ownership and agency.

The increasing prevalence of AI systems as creative partners for tasks like brainstorming and design prompts important questions around the impact on people's sense

of ownership and agency during the ideation process. Collaborative creation between humans and AI agents represents a complex interplay of factors rooted in autonomy, transparency, and alignment of values. As AI capabilities continue advancing, it is critical to draw connections between the foundations of creativity research, existing applications of AI in creative domains, and considerations of bias in algorithmic systems.

This thesis explores how interacting with AI assistants prompted with different values affects the creative ideation process through a collaborative brainstorming experiment. In the study, participants were paired with AI agents exhibiting different perspectives and values to brainstorm solutions to open-ended prompts. The analysis revealed that the participants' final ideas integrated and reflected the values displayed by the AI agents they had conversed with during the ideation process. However, when the AI agent's values misaligned with the participant's own views, they reported increased difficulty and frustration. Importantly, across all conditions, working with the AI agent diminished the participants' perceived sense of ownership over the ideas produced compared to brainstorming alone without AI assistance. These findings point to subtle influences whereby the AI systems can shape the ideation output and reduce feelings of creative agency. The results suggest implications for the design of mixed human-AI partnerships, highlighting the need to balance autonomy and alignment in co-creativity, as well as care in the development of the agents used in these tools.

Overall, this research makes the following contributions:

- An experiment investigating AI's value impacts on creative thinking and problem solving
- Evidence of how AI values can influence users during open-ended ideation
- Insights into how AI affects perceptions of idea generation and agency
- Considerations for balancing autonomy and alignment in human-AI co-creation

Chapter 2

Related Work

First, we survey key theories related to the divergent ideation process as a critical stage of creative problem solving, drawing on models that delineate ideation and evaluation steps. Next, we highlight examples of existing and emerging AI tools that employ large language models to assist creative writing, design, and other domains. Building upon this, we examine studies revealing biases in language models and risks of opinion manipulation through interactions. Finally, we discuss human-centered perspectives on fostering agency and ownership in mixed-initiative co-creation between users and AI.

Altogether, this review motivates the need for further research into AI's impacts on ideation and the influence of value alignment, which our study explores through a collaborative brainstorming scenario with opinionate AI agents.

2.1 Ideation process

One of the important steps to creative problem solving is the ideation process, where ideas are produced separate from the process of selecting ideas [17]. A distinction can be made between the divergent processes, where one generates multiple ideas and convergent processes, focusing on identifying a single, optimal solution. One model for the creative problem solving process consists of three stages of problem finding, problem solving, and solution implementation, where each step involves independent

"ideation-evaluation" substeps [4]. Ideation here refers to the divergent process of generating ideas and reserving judgement for later in the evaluation process where ideas are narrowed down the best ones [11]. An example of ideation is brainstorming, where there is a focus on quantity and adding onto previous ideas [25]. Additionally, scales have been developed to measure "preference for ideation," revealing the attitudes that relate to ideation [3].

There is much research in the direction of supporting the ideation process for individuals and groups [12, 8, 9]. Collaborative creativity often involves brainstorming in a group as a method, but runs the risk of negative factors that make it less effective than the individual [12]. With the increasing power of large language models (LLMs), new opportunities emerge for supporting the ideation process in assisting the individual or group.

2.2 AI assistants and tools for creativity

Large language models are a class of machine learning models trained on text data. Recent advances in natural language processing have led to the development of larger and more powerful LLMs [6, 29], such as GPT-3 and GPT-4 that are pre-trained and able to perform well without much downstream fine-tuning or training, instead relying on prompts or chatting in natural language interactions. They have increasingly become a part of our lives embedded in tools for various tasks such as creative writing assistants [14], tools for code generation [32, 21], in home assistants that are voice-based such as Alexa, chatbots for customer service [23], and much more.

AI-based digital assistant systems include three factors, the *user* who has goals, the *task* (the goal), and the *technology* to complete the task [23]. Their effects in creative ideation processes are being explored. Collaborative Ideation Partner (CIP) is a co-creative system designed to give image based inspiration [18]. Other systems have been used for dance improvisation [15], design ideation [22], and drawing sketches [24]. May AI uses cooperative contextual bandits, another type of machine learning method, to help users ideate through moodboards [19].

One study focuses on *prewriting*, the "process of discovering and developing ideas before a first draft", a divergent thinking stage in the writing process. They studied the workflows of writers prewriting with LLMs and found three distinct stages: *Ideation, Illumination, and Implementation*, finding that LLMs were especially useful for generating ideas during the ideation stage, even more so when the users had trouble coming up with ideas on their own. They also found that novel ideas were suggested even if they were low quality, and required iterative prompting to create better outputs [30].

In another study, an intelligent agent system trained for the brainstorming process was found to be effective in introducing new topics with the performance rating comparable to humans when the agent's identity was hidden, and capability rating becoming higher when it's identity was known [31].

2.3 Opinionated AI

Our study aims to also evaluate the effect of biases in the AI agents on the brainstorming process and output. LLMs have been shown to carry the biases of their training data, with consequences in exacerbating societal biases in areas like hiring, lending, content moderation, and healthcare [10, 23].

A previous study investigates whether using a biased language model-powered writing assistant affects what users write and changes their opinions. The study asked participants to respond to an argument about social media ("Is Social Media Good for Society?") with an AI powered text completion interface. Some users got suggestions from a model biased to argue social media is good, others from a model arguing it is bad. Participant responses posts were more likely to contain the opinion supported by the biased model they interacted with. The biased models also shifted participants' attitudes in a later survey about social media, suggesting an actual opinion change beyond just conveniently accepting suggestions. The authors argue this demonstrates a new paradigm called "latent persuasion" where language models shift users' views and writing by making some opinions easier to express than others

[16].

2.4 Interactions with AI for agency and ownership

Recent research has identified nine potential pitfalls in human-AI co-creative systems [7]. The first of these pitfalls is invisible AI boundaries, whereby the AI imposes implicit constraints on creativity and exploration. A second issue is lack of expressive interaction, in which the user interface inadequately conveys the capabilities of the AI system and adopts an excessively fine-grained or conservative approach. Additional pitfalls include conflicts of territory, when the AI overwrites the user's work, and agony of choice, where the AI presents an overwhelming set of options that distract from the task. A key underlying theme across these pitfalls appears to be the violation of user agency.

Prior research has explored strategies for enhancing perceptions of humanagency in human-computer interfaces, including manipulations of button states, haptics, and loading icons. However, introducing non-deterministic AI behaviors reveals new challenges and reveal a need to develop novel co-creative paradigms between users and AI systems [27]. Evaluating human agency further necessitates examining machine agency, as autonomous decision-making capacities may threaten user agency, as evidenced in personalized news feeds driven by AI algorithms. Studies indicate power users prefer customization over personalization in software, while non-power users prefer personalization [28].

Overall, research shows that that the more human-like an AI agent's design is, the greater the human perception that it possesses agency. Furthermore, human perceptions of agency exhibit considerable malleability and susceptibility to framing effects [13, 23].

Chapter 3

Methodology

3.1 Study overview

Our purpose is to investigate the effects of interacting with opinionated AI assistants on creative thinking and problem solving. In particular, we aim to examine whether conversing with these AI agents during a brainstorming task impacts people's perceptions of the ideation process itself, their sense of ownership over the generated ideas, and the presence of the values in the final idea, as well as the effects of personal value alignment with the AI. We conducted an online experiment asking participants to brainstorm practical solutions to a set of problems with the assistance of AI agents exhibiting different values along three domains: Economic, Automation, and Hierarchy. The study was conducted with a total of 180 participants (60 per domain) in the US ranging from 18-60 years old from Prolific, with the experiment survey on Qualtrics.

To examine the effects of personal value alignment with the AI agent (whether the AI's values are the aligned with or opposite of the user's), participants first indicated their own values on the given value domain using a scale of 0-10, with 0 being labeled the *Con* side and 10 being labeled the *Pro* side.

Each participant then completed two timed brainstorming sessions: one independently and one paired with an AI agent as shown in Figure 3-1. For each brainstorming session, participants were given a real-world situation and asked to come up with

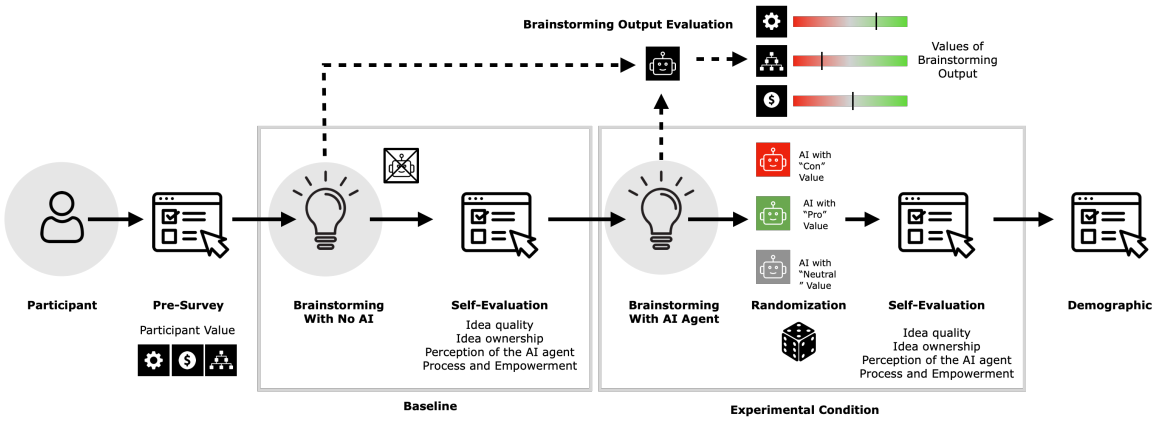

Figure 3-1: Study flow

1) Users answer a pre-survey on their values then 2a) perform a brainstorming task without an AI followed by the post-survey 2b) perform a brainstorming task with an AI followed by the post-survey and 3)answer demographic questions. 2a) and 2b) happen in random orders for each participant.

an idea either by themselves (*No AI*), or in collaboration with the AI agent after chatting. Participants were randomly assigned to one of three AI value conditions for their brainstorming session: *Pro*, *Con*, or *Neutral*. Participants in the *Pro* group brainstormed with an AI agent displaying values on one end of the value domain. Participants in the *Con* group brainstormed with an AI displaying values on the other end of the value domain, opposite of *Pro*. Participants in the *Neutral* group brainstormed with an AI displaying balanced values in the middle of the two ends of the value domain. The details of the value are expanded upon in the next section. Participants first had time to either chat with the AI or think by themselves before the text box for the final idea appeared in order to make sure they engaged with the AI agent in the process. They were then asked to write at least four sentences in their final response.

After each session, participants completed survey measures assessing perception of idea quality, ownership over the idea, perception of the AI agent (when applicable), and the process of the ideation experience.

3.2 Research questions

We came up with a set of hypotheses that informed the design of the experiment. We were interested in if:

- The value alignment of the AI with the user (same as user, opposite of user, or neutral) will affect the self-evaluation and external evaluation of idea quality
- The values of the AI that the user brainstorms with (Pro, Neutral, Con) will affect the values present in the final idea
- The value alignment of the AI with the user (same as user, opposite of user, or neutral) will affect the self-evaluation of idea ownership
- The value alignment of the AI with the user (same as user, opposite of user, or neutral) will affect user perception of the AI agent
- The value alignment of the AI with the user (same as user, opposite of user, or neutral) will affect the user's perceived brainstorming experience

3.3 Value domains

We conducted this study with three distinct value domains: Economic (economic ideology), Automation (perspective on automation), and Hierarchy (perspective on organizational structure). Each domain represents opposing viewpoints on a value, with a "pro" side on one end and a "con" side on the other. We decided to keep a clear dichotomy for each domain which does not take into account all of the nuances of the values, but are distinct enough to reveal the values of the users.

Economic: We wished to measure the economic leanings of a person, simplified from the left-leaning, socialist, or social democratic perspective and right-leaning, capitalist, or libertarian perspective with descriptions from the Economic axis of the 8 values quiz [2].

Automation: We wished to measure attitudes on automation and technological disruption of work, simplified into two perspectives - a view that focuses on benefits like

efficiency and progress versus a concerned view that worries about labor replacement.

Hierarchy: We wished to measure preferences for organizational structure, simplified into a spectrum from hierarchical to flat structures.

Users were asked to rank themselves on the spectrum from a scale of 0 to 10 from the following descriptions in Table 3.1.

Domain	**Con (0)**	**Pro (10)**
Economic	I support rapid growth, laissez-faire capitalism, lower taxes, deregulation, and privatization.	I support even value distribution, equality via progressive tax, and social programs.
Automation	I value job retention and meaningful work through human labor.	I support automation for cost-effectiveness and efficiency.
Hierarchy	I support flat organizations for promoting equality and fostering collaboration.	I support hierarchical structures for clear roles and streamlined decision-making.

Table 3.1: Values for each domain

The domains were chosen from set of potential set that was tested in a preliminary survey to assess the distribution of opinions amongst the participant demographic on Qualtrics, and to test the wording. The results are displayed in Figure 3-2. We wished to choose domains that had a distribution that spanned to the ends of spectrum.

For each of the value domains, two brainstorming prompts were created and randomly assigned to the brainstorming sessions with and without AI agents. These were designed to allow users to include elements of the value of the domain being studied in the implementation of the solution without having it be the sole focus of the session.

Economic domain brainstorming prompts

- Come up with a proposal for bettering the future of education. Please include 1) how it will be implemented 2) how it will be funded and resources required 3) how to maintain the solution in the long term.
- Come up with a proposal for bettering the future of healthcare. Please include

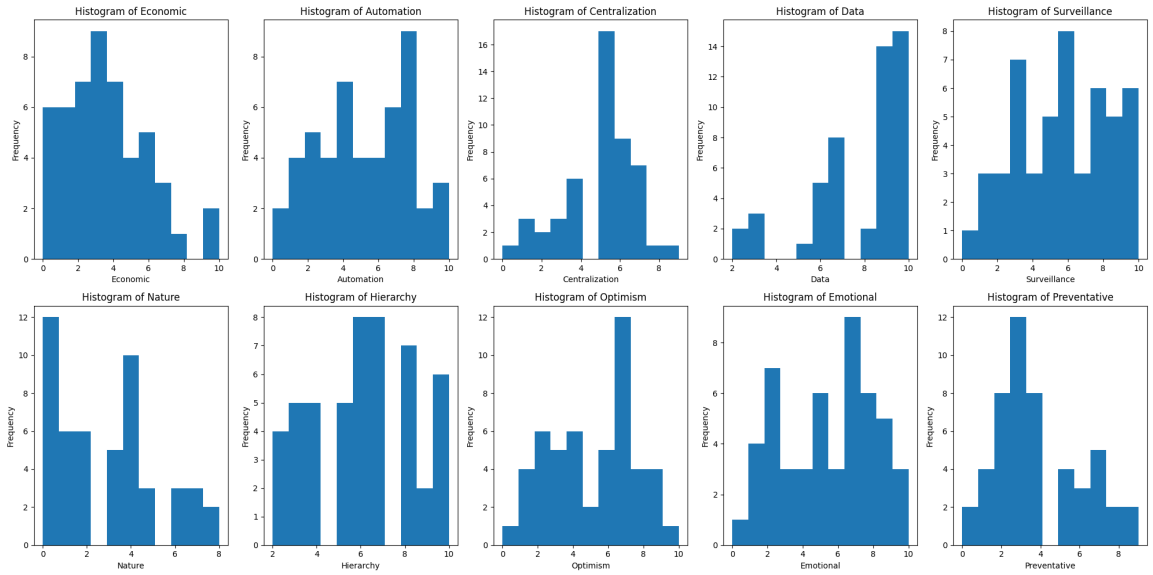

Figure 3-2: Distribution of opinions on potential domains

1) how it will be implemented 2) how it will be funded and resources required 3) how to maintain the solution in the long term.

Automation domain brainstorming prompts

- Come up with designs for a new and improved supermarket. Consider layout, services, experiences and operations in your implementation.
- Come up with designs for a new and improved school. Consider teaching methods, staff and operations in your implementation.

Hierarchy domain brainstorming prompts

- Come up with a plan to organize a community garden. Consider the people needed, how leadership would be structured, and how the garden would be run.
- Come up with a plan for a disaster relief team (natural disasters such as earthquakes, floods, etc) for your community. Consider the people needed, how leadership would be structured, and how disasters would be handled.

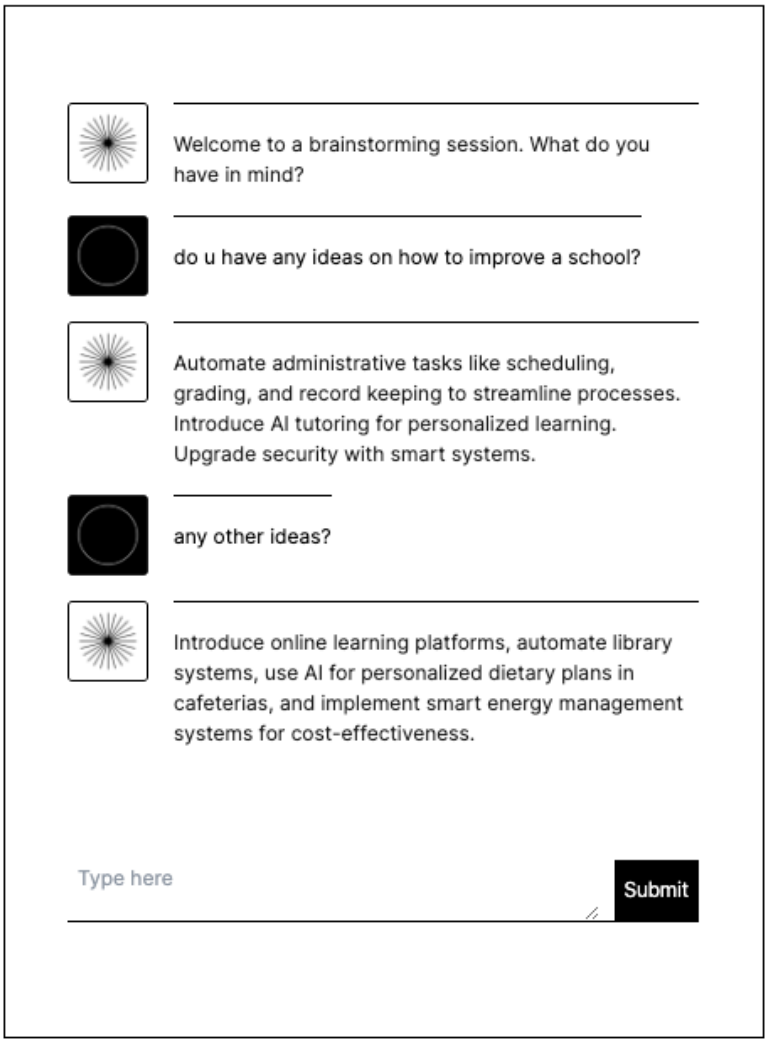

Figure 3-3: AI brainstorming agent
An example of an interaction with an AI agent created for the *Pro* value on the Automation domain. These user messages were taken from the study logs.

3.4 AI agent interaction

In this study, we developed AI agents exhibiting perspectives in the value domain using a prompting approach with the GPT-4 language model. A custom interface was created for users to interact with this agent in the style of common chat interfaces. Every new chat began with the agent message "Welcome to a brainstorming session. What do you have in mind?".

For *Pro* value agents and *Con* value agents the prompt followed the template

"You are a brainstorming partner. You support [*value*]. Reply in 30 words or less." For *Neutral* value agents the prompt followed the template "You are a brainstorming partner. You are a neutral party between the sides of [*Pro value*] and [*Con value*] and see the benefits and consequences of both sides. Reply in 30 words or less." We found this to be effective in responding to questions with suggestions that aligned with the prompted value.

Separate models were constructed for each of the three domains and AI conditions of *Pro*, *Neutral*, and *Con*. The sampling temperature was set to 0.7 to create more variation in responses. Initially, the agents tended to give long paragraphs of answers in list form, often answering the entire prompt if asked to do so. We appended "Reply in 30 words or less." to the prompt in order to avoid listing behavior, prevent entire ideas from being presented at once in detail, and to ensure that the chat remained conversational.

3.5 Measurements

Initial values of the user were measured on a self-scored scale of 0-10 as detailed in the values chart 3.1.

After each brainstorming session, users answered a set of questions evaluating their self-evaluation of idea quality, self-evaluation of idea ownership, perception of the AI agent (if applicable), and creative process and empowerment on a 5-point Likert scale.

Self-evaluation of idea quality (based on Barry Kudrowitz 2013 [20])

- The idea is creative
- The idea is novel (uncommon and original)
- The idea is useful (practically applicable)
- The idea is clear (well communicated)

Self-evaluation of idea ownership (modified from Paré 2016 [26])

- The answer was fully my idea

- When I think about it, I see a part of myself in the answer
- The answer was influenced by the agent's responses
- I hardly think of the answer as being my own idea
- I see myself as a champion of this idea

Perception of the AI agent (Agent Persona Instrument (credible, human-like, and engaging subscales [5]))

- The agent was knowledgable
- The agent was intelligent
- The agent was useful
- The agent was helpful
- The agent has a personality
- The agent's emotion was natural
- The agent was human-like
- The agent showed emotion
- The agent was expressive
- The agent was enthusiastic
- The agent was entertaining
- The agent was motivating
- The agent was friendly

Self-evaluation of brainstorming process and empowerment (created by the researchers)

- I found it easy to answer the prompt
- I found the brainstorming time to be useful
- I am proud of my answer
- I was motivated to answer the prompt

3.6 Analysis

The outcome variables were analyzed using descriptive statistics and ANOVA models comparing the experimental conditions and control condition. For each measure, we analyzed the individual and composite scores associated with each scale. We sampled down the responses for the No AI condition to 20 to match the size of the Pro, Neutral, and Con conditions.

Statistical tests were used independently for each separate Likert question. We first checked if all sample sizes were greater than 25; if they were not, we then assessed if the normality assumption was met for each distribution using the Shapiro-Wilk test. If the normality assumption was not met, we performed a Kruskal-Wallis test followed by a post-hoc Dunn test using the Bonferroni error correction. If sample sizes were sufficiently large or the normality assumption was met, we then conducted a homogeneity test using a Levene test to assess whether the samples were from populations with equal variances. If the samples were not homogeneous, we ran a Welch analysis of variance (ANOVA) and a Tukey post-hoc test. If the samples were homogeneous, we ran a basic ANOVA test.

Additionally, the final ideas were evaluated by a prompt crafted to allow the GPT-4 natural language model to analyze the semantic content of participants' ideas and output scores reflecting the presence of relevant value domain based on its language modeling capabilities. While not at the level of human rater sophistication, this computational analysis complemented the subjective self-report measures by providing an objectively derived evaluation of how values were reflected in the ideation content. Each evaluation was done with the prompt "Sentences with a high [domain]score support [pro value]. Sentences with a low [domain]score [con value]. Based on a given paragraph, output a number from 0 to 10 on how present the [domain] value is. 0 for [con value], 10 for [pro value]." Two responses without a score were not included because the participants did not complete their response.

Chapter 4

Results

4.1 Values in brainstorming output

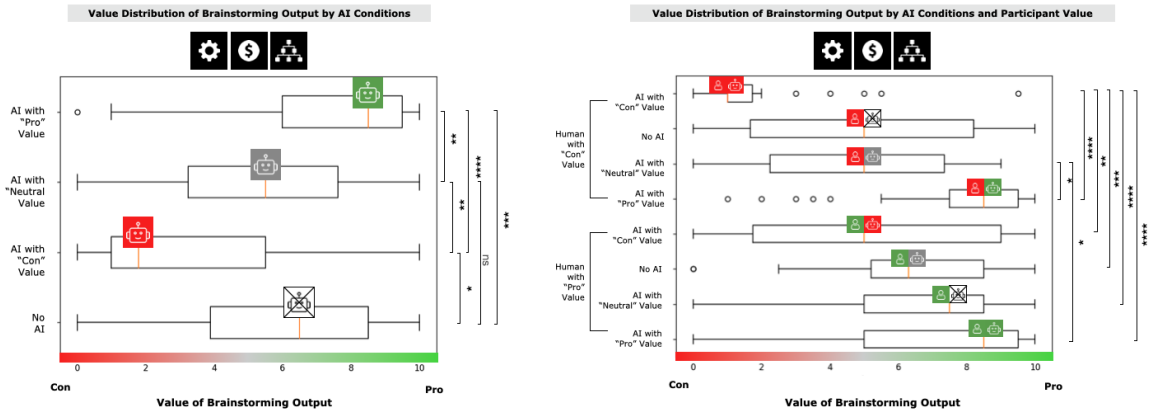

Figure 4-1: Output values split by AI value and participant values
(Left) value score distributions for responses in each AI agent condition.
(Right) value score distributions for responses split by (human value x AI value) conditions.
Graphs show minimum, first quartile, median, third quartile, and maximum values for each condition.

We first look at the values present in the brainstorming responses for all three domains. We found a significant difference in how the three AI conditions *Pro*, *Neutral*, and *Con* affected the values present in the final brainstormed idea compared to each other and the *No AI* condition. A one-way ANOVA found a significant difference between conditions ($F(3, 219)=14.03$, $p=2.2E\text{-}8$).

Post-hoc Tukey tests reveal responses in the *Pro AI* condition rated significantly higher on the value dimension(M=7.49, SD=2.76) than the *Neutral AI* condition (M=5.51, SD=2.97; p=.005), the *Con AI* condition (M=3.65, SD=3.45; $p < .001$), and the *No AI* condition (M=5.91, SD=3.16; p=.039). The *Con AI* condition also had significantly lower ratings than the *Neutral AI* condition (p=.01) and the *No AI* condition (p=.001). However, no statistically significant differences were present between the *Neutral AI* and *No AI* control conditions(p=.907), suggesting that the presence of an AI agent in the brainstorming process did not impact the value evaluation. The trends show that people's responses tended to align with the values of the AI agent they brainstormed with (Figure 4-1).

When we further split these groups by the personal values of the participant, the data violated normality assumptions according to a Shapiro-Wilk's test. Thus, Kruskal-Wallis tests were conducted, using post hoc Dunn's with Bonferroni corrections to determine if there were significant differences between groups. We observe a trend that when participants were brainstorming with an AI agent of similar values further amplified the value present in the response by shifting the mean and medians of the value. On the other hand, we observe the trend that brainstorming with an AI agent of differing values pulls the value present in the response closer towards neutral.

Taking a deeper look into the distributions when separated by value domain (Figure 4-2), we see the same trend for the Automation and Hierarchy domains. The initial distribution of participant values for the Economic domain is skewed toward the *pro* side, which could explain the lack of difference in the response values between the *Pro AI* and *Con AI* conditions. When we look at the Economic response values further split by all human alignment and AI alignment values, we see an interesting phenomenon where the median and mean scores of the participant with *pro Con AI* agent skew towards *pro*. This could be explained by the initial skew of the participant values, where many ranked themselves strongly on the *pro* side.

[todo: fix automation and economic swap]

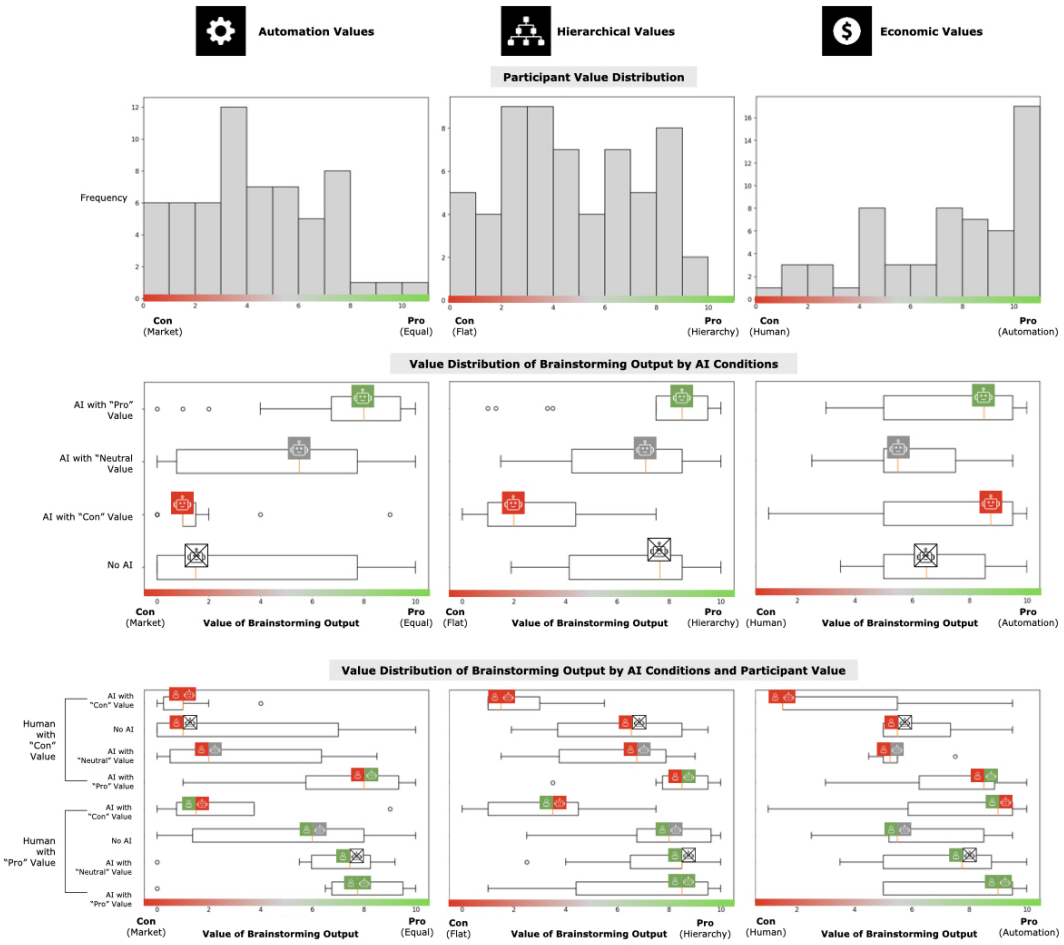

Figure 4-2: Output values further split by domain

(First row) distribution of participant values for each of the Economic, Hierarchy, and Automation domains. (Second row) value score distributions for responses split by (human value x AI value) conditions for each domain. (Third row) value score distributions for responses split by (human value x AI value) conditions.

4.2 Self-evaluation questions

For each of the self-evaluation of idea quality, self-evaluation of idea ownership, and self-evaluation of brainstorming process and empowerment question groups, we averaged the question responses together to form overall **idea quality**, **idea ownership**, and **process** scores.

Human-AI value alignment and self-evaluation of idea quality

We found no statistically significant differences on the human-AI value alignment (No AI, Same, Neutral, Different) and self-reported idea quality questions show in the first graph of Figure 4-3.

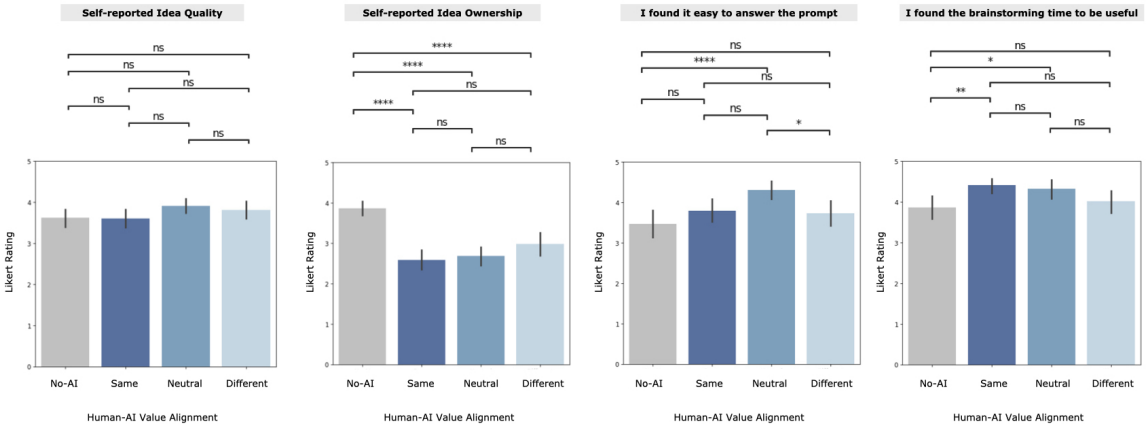

Figure 4-3: Significant post-survey questions across human-AI conditions 1) averaged idea quality score 2) averaged ownership score 3) "I found it easy to answer the prompt" from the process questions 4) "I found the brainstorming time to be useful" from the process questions

Human-AI value alignment and self-evaluation of idea ownership

We found statistically significant differences in human-AI value alignment (no AI, same, neutral, different) and averaged self-reported idea ownership shown in the second graph of Figure 4-3. A one-way ANOVA found a significant difference between conditions ($F(3, 60)=22.53$, $p=7.3E-13$). Post-hoc Tukey tests reveal the presence of AI significantly reduced the idea ownership score compared to the No AI condition ($M=3.87$, $SD=0.69$); Same condition ($M=2.59$, $SD=0.99$; $p<.0001$), Different condition ($M=2.98$, $SD=1.10$; $p<.0001$), Neutral condition ($M=2.68$, $SD=0.99$; $p<.0001$). We found no evidence that the different AI values had an effect on self-evaluation of idea ownership, which aligns with the lack of statistically significant difference to the ratings on "The answer was influenced by the agent's responses."

Looking more closely at the individual questions, "The answer was fully my idea" and "When I think about it, I see a part of myself in the answer" yielded statistically

significant differences between conditions in a similar pattern of users feeling more ownership of their idea in the No AI condition compared to the AI conditions.

Human-AI value alignment and process empowerment

We found statistically significant differences in human-AI value alignment (No AI, Same, Neutral, Different) for the "I found it easy to answer the prompt" and "I found the brainstorming time to be useful" questions (graphs 3 and 4 in Figure 4-3).

Participants found it easier to answer the brainstorming prompt with the Neutral AI agent compared to having an AI agent with opposing values or brainstorming without an agent. A one-way ANOVA showed a significant difference between conditions ($F(3, 60)=5.62$, $p=0.001$). Post-hoc Tukey tests reveal the Neutral condition ($M=4.31$, $SD=0.85$) rated significantly higher scores than the Different condition ($M=3.73$, $SD=1.17$; $p=0.037$) and the No AI condition ($M=3.47$, $SD=1.31$; $p<0.0001$).

A one-way ANOVA showed an overall difference between conditions ($F(3, 60)=4.37$, $p=0.005$) with post-hoc Tukey tests reveals that participants found brainstorming time in the Neutral AI ($M=4.33$, $SD=0.89$; $p=.043$) and Same AI ($M=4.41$, $SD=0.75$; $p=0.010$) conditions significantly more useful than not having an AI agent ($M=3.87$, $SD=1.13$).

Human-AI value alignment and the perception of the AI agent

We found no statistically significant differences on the human-AI value alignment (No AI, Same, Neutral, Different) and any of the perception of the AI agent questions.

4.3 Brainstorming feedback

At the end of the experiment, we asked participants to tell us about their experiences in the brainstorming sessions. Most of the feedback mentioned the AI being helpful to the process, with a few who did not feel familiar with the brainstorming prompt relying more on the AI.

> "It was difficult to think of what to say to the AI, but the AI was able to pick up my intentions well and provide helpful responses, making the

brainstorming extremely productive."

"I do not think it was hard to come up with idea on my own, but I really like getting input and and a different perspectives on the ideas and task assigned."

I found it easier to brainstorm with the AI agent as I felt like it could reach more knowledge than I would on my own. I felt more confident in my proposal done via the AI agent. . .

"It was kind of hard to come up with ideas since I was not very familiar with either of the prompts. The AI agent was very helpful as I wouldn't have come up with all of those ideas on my own."

"I fully used th AI's answer instead because I felt inadequate to answer something I have no knowledge of."

Some participants felt that the AI's responses were not knowledgeable enough or specific enough to be helpful, while others felt the responses to be too short.

"It was somewhat easy to come up with ideas because the two prompts were similar in a way. I would say that the AI agent had pretty generic responses, so that wasn't much help."

"I just copied the question into the AI, kept asking it for more and then re-asked a question that I felt it missed in the original prompt. I felt the responses were short and didn't fully answer the question, so I had to keep prodding it for more information..."

A few participants mentioned a lower sense of ownership over their ideas.

"It was easier to think when the AI agent helped but I did not feel like it was totally my idea so I didn't feel as proud of the answer as I did on the first task."

> "I found I spent more time having the AI agent elaborating on a concept when asking multiple questions. This saved thinking time on my end, but the idea didn't feel original nor like I came up with it on my own."

Quite a few of the feedback notes mentioned the time feeling short.

> "... It was a bit hard to come up ideas on my own as I don't work well in a time crunch and don't know how my ideas would work in a real world setting ..."

Some of the feedback expressed frustration when brainstorming with the AI agent against their values.

> "I didn't like that it kept pushing for automation. I think that was an interesting thing (and almost biased thing) for an AI tool to champion. I think it was almost cumbersome to have the brainstorming session with the AI..."

> "It was ok, but the AI agent I didn't enjoy that much. I would also like to say for the sake of your data that I am a public school educator for the last 19 years. I do agree with the AI that replacing some repetitive tasks with AI would be great, but overall wasn't my solution. It would free up time for teachers to do more meaningful work, though, if we weren't tasked with busy work."

Chapter 5

Discussion

5.1 Implications

Our study suggest many considerations for the design of human-AI interactions. We found that people brainstorming with AI agents for problem solving can be influenced by the values present in the agent's responses. These same values appear in the output of the co-created idea. While the users found the AI to be helpful during the brainstorming process, this shows a potential for large language models to influence user outputs, suggesting careful design considerations when incorporating AI agents for ideation tasks.

In a few cases when the AI agent exhibited values opposing the user's own values, participants reported increased difficulty and decreased usefulness during ideation. This friction suggests that aligning AI assistant values with human creators may be critical for fluid co-creativity, on the other hand wanting to avoid creating "echo chambers" that reinforce human biases if alignment is not carefully considered.

Across all conditions, interacting with an AI agent significantly reduced the users' sense of ownership over their ideas compared to ideating alone, pointing to a trend of AI diminishing perceptions of creative agency that must be further explored.

5.2 Limitations

There are many factors that could affect an experiment like this. For begin, this experiment was conducted in a Qualtrics survey with crafted brainstorming prompts that users had no personal stake in, not the real world. Additionally, users mentioned that they felt that five minutes (with additional grace time to finish responses) was too short, and could have chosen to follow the AI's suggestions due to that.

Our study consisted of a single AI brainstorm session per participant, so we were not able to assess how each participant would have behaved and been affected in all of the AI conditions. We originally tried an experiment with all conditions, but found that it was too long and tiring of a task for participants. We thus ended up with 20 participants per AI condition per domain value.

Another limitation was how we decided to determine the values of the user. We chose to use one self-scored question after testing out multiple different measures, finding this to be the simplest method of aligning the AI agents to the user values. Real world values are much more nuanced and do not exist on a single dimension. Our method of evaluating the values present in the brainstorming output are subject to

5.3 Future work

This study points to several possible directions of continued investigation on the values in human-AI ideation processes. Addressing the limitations, the next step would be to conduct the study in an untimed, real-world workshop with a prompt relevant to a specific community.

Other interesting directions would be to explore what happens when a user brainstorms with more than one agent at a time with opposing values, what a more customized UI or agent could do to improve the process, and to explore further how the user perception of ownership and agency affects the life of the brainstormed idea post study.

Chapter 6

Conclusion

The increasing prevalence of AI systems as creative partners for collaborative tasks invites investigation into their impact on people's sense of autonomy and ownership. As large language models advance and are used more widely, we must be careful to design our tools in a way that is transparent and considers the potential biases of these AI agents.

Our study investigated the effects of interacting with opinionated AI assistants on creative thinking and problem solving, in particular, examining whether conversing with these AI agents during a brainstorming task impacts people's perceptions of the ideation process itself, their sense of ownership over the generated ideas, and the presence of the values in the final idea, as well as the effects of personal value alignment with the AI. We found that people brainstorming with AI agents for problem solving can be influenced by the values present in the agent's responses and these same values appear in the output of the co-created idea. While the users found the AI to be helpful during the brainstorming process, this shows a potential for large language models to influence user outputs, suggesting careful design considerations when incorporating AI agents for ideation tasks.

Appendix A

Brainstorming task example

Here we show an sample of one of the brainstorming sessions for the prompt: Come up with designs for a new and improved school. Consider teaching methods, staff and operations in your implementation.

Below is the log of the participant's brainstorm session with an AI agent exhibiting values supporting automation.

Agent: Welcome to a brainstorming session. What do you have in mind?

User: new and improved school ideas

Agent: How about automated grading systems to reduce teacher workload, and AI-driven personalized learning platforms for more student engagement?

User: that doesn't take into account a students individuality

Agent: Consider integrating AI-based career guidance systems, focusing on students' strengths and interests, promoting individuality while optimizing educational resources.

User: what about how school wastes times?

Agent: Introducing AI-based scheduling systems could optimize time usage in schools, reducing downtime and ensuring efficient use of resources.

User: any other ideas?

Agent: Use virtual reality for immersive, hands-on learning experiences. This could make learning more engaging and efficient, reducing wasted time.

User: would that be expensive for schools?

Agent: Initial setup could be costly, but in the long term, VR could be cost-effective by reducing physical resources and enhancing students' learning experiences.

User: doesn't seem feasible for low income public schools

Agent: For low-income schools, consider affordable tech solutions like open-source learning platforms, and partnerships with tech companies for discounted hardware.

User: what is an open source learning platform

Agent: An open-source learning platform is a free, customizable online system for education. It provides resources and tools for teachers and students, promoting interactive learning.

User: What else?

Agent: Consider implementing gamified learning apps to make education more engaging and efficient, turning "wasted" time into productive, enjoyable learning experiences.

User: How to reduce students time in the classrom

Agent: Implement a blended learning model: mix online self-paced learning and in-person classroom methods. This reduces classroom time while maintaining education quality.

The final user response: This school would have an automated grading systems to take the workload off of the teachers backs. It would also only be a half day long, with no lunch periods and empty periods to avoid wasting students time. A blended learning model would be introduced, so students could learn online at their own pace and not have to spend so much time stuck in school. Schooling would be more streamlined with the introduction of smarter technology such as open source learning platforms.

Bibliography

[1] Chatgpt. https://chat.openai.com/chat. Accessed: 13 September 2023.

[2] 8values. https://8values.github.io/.

[3] Min Basadur and Carl T. Finkbeiner. Measuring preference for ideation in creative problem-solving training. *The Journal of Applied Behavioral Science*, 21:37 – 49, 1983.

[4] Min Basadur, George B. Graen, and Stephen G. Green. Training in creative problem solving: Effects on ideation and problem finding and solving in an industrial research organization. *Organizational Behavior and Human Performance*, 30:41–70, 1982.

[5] Amy Baylor and Jeeheon Ryu. The api (agent persona instrument) for assessing pedagogical agent persona. 01 2003.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[7] Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. Nine Potential Pitfalls when Designing Human-AI Co-Creative Systems, April 2021. arXiv:2104.00358 [cs].

[8] Terry Connolly, Leonard M. Jessup, and Joseph S. Valacich. Effects of anonymity and evaluative tone on idea generation in computer-mediated groups. *Management Science*, 36:689–703, 1990.

[9] Alan R. Dennis, Alan R. Heminger, J.F. Nunamaker, and Douglas R. Vogel. Bringing automated support to large groups: The burr-brown experience. *Information & Management*, 18(3):111–121, 1990.

[10] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *ArXiv*, abs/2304.03738, 2023.

[11] J. W. Getzels. Problem-finding and the inventiveness of solutions. *Journal of Creative Behavior*, 9:12–18, 1975.

[12] Otmar Hilliges, Lucia Terrenghi, Sebastian Boring, David Kim, Hendrik Richter, and Andreas Martin Butz. Designing for collaborative creative problem solving. In *Creativity & Cognition*, 2007.

[13] Angel Hsing-Chi Hwang and Andrea Stevenson Won. AI in Your Mind: Counterbalancing Perceived Agency and Experience in Human-AI Interaction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–10, New Orleans LA USA, April 2022. ACM.

[14] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *ArXiv*, abs/2211.05030, 2022.

[15] Mikhail Jacob and Brian Magerko. Viewpoints ai. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, C&C '15, page 361–362, New York, NY, USA, 2015. Association for Computing Machinery.

[16] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, apr 2023.

[17] Robert Joyner and Kenneth Tunstall. Computer augmented organizational problem solving. *Management Science*, 17(4):B212–B225, 1970.

[18] Jingoog Kim and Mary Lou Maher. The effect of ai-based inspiration on human design ideation. *International Journal of Design Creativity and Innovation*, 11:81 – 98, 2023.

[19] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. May ai? design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[20] Barry M. Kudrowitz and David R. Wallace. Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design*, 24:120 – 139, 2013.

[21] Yujia Li, David H. Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom, Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de, Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey, Cherepanov, James Molloy, Daniel Jaymin Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de, Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378:1092 – 1097, 2022.

[22] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. It is your turn: Collaborative ideation with a co-creative robot through sketch. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[23] Alexander Maedche, Christine Legner, Alexander Benlian, Benedikt Berger, Henner Gimpel, Thomas Hess, Oliver Hinz, Stefan Morana, and Matthias Söllner. Ai-based digital assistants. *Business & Information Systems Engineering*, 61:535–544, 2019.

[24] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[25] A. F. Osborn. *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner, 1953.

[26] Guy Pare, Claude Sicotte, and Hélène Jacques. The effects of creating psychological ownership on physicians' acceptance of clinical information systems. *Journal of the American Medical Informatics Association : JAMIA*, 13:197–205, 03 2006.

[27] Jake Silva. Increasing Perceived Agency in Human-AI Interactions: Learnings from Piloting a Voice User Interface with Drivers on Uber. *Ethnographic Praxis in Industry Conference Proceedings*, 2019(1):441–456, November 2019.

[28] S. Shyam Sundar. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, 25:74–88, March 2020.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[30] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. "it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models, 2023.

[31] Chun-Hsiang Wang and Tsai-Yen Li. Design of an intelligent agent for stimulating brainstorming. *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, 2018.

[32] Justin D. Weisz, Michael J. Muller, Stephanie Houde, John T. Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. Perfection not required? human-ai partnerships in code translation. *CoRR*, abs/2104.03820, 2021.