Received: 18 August 2025 | Revised: 15 January 2026 | Accepted: 6 May 2026 | Published online: 28 May 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–22
DOI: 10.47852/bonviewAIA62027260

BON VIEW PUBLISHING

Future AI Systems: Human–AI Collaborative Teams Outperform Human-Only Teams in Design Ideation

Joseph M. Makokha[1,*]

[1] *Department of Informatics, University of Oslo, Norway*

Abstract: This paper explores how an artificial intelligence (AI) system acting as a collaborative partner can enhance the concept development stage in engineering design and other ideation contexts. We introduce the concept of a "Disruptive Interjector" (DI), an AI system that monitors user actions and offers suggestions based on those actions. We demonstrate its application in a human-subject study focusing on collaborative teamwork. Specifically, we compare the performance of teams comprising two humans (HH) with teams consisting of a human and an AI (HAI) in a design sketching task based on a given prompt. The results indicate that HAI teams consistently generated ideas that received higher ratings from trained human judges compared to HH teams. Subsequently, this paper makes three significant contributions to AI research and practice. First, it introduces the DI, a conceptual model of an AI designed for divergent idea stimulation. Second, it provides insights into design considerations for creating such AI systems, particularly as AI becomes increasingly integrated into various fields like engineering design. Finally, it offers a replicable method for conducting studies involving AI and provides empirical evidence that AI can outperform humans as collaborative partners in concept development tasks.

Keywords: generative AI, human–AI collaboration, human–AI creativity

1. Introduction

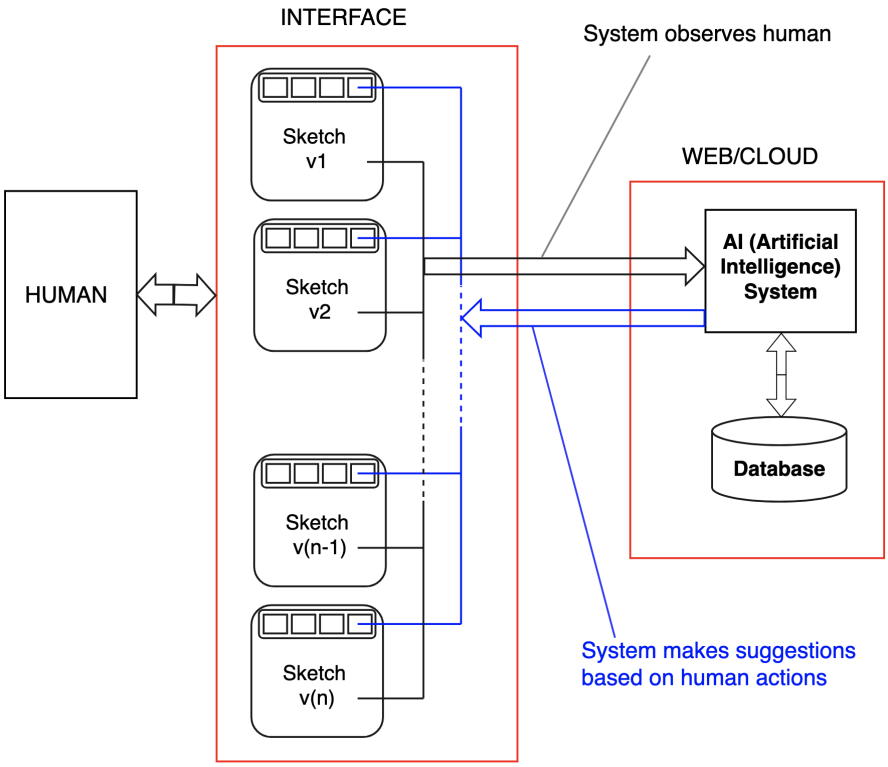
The past few decades have witnessed the emergence of a class of computer systems that implement artificial intelligence (AI) with capabilities approaching humans in a variety of domains such as detecting fractures and tumors in medicine [1–3], automating tasks in data science, generating art [4], and various kinds of robotics and autonomous cars [5], among others. As a result of this proliferation of AI systems, people are frequently collaborating [6] with AI in human–AI (HAI) teams spanning many different contexts. There is therefore a need for researchers, practitioners, decision-makers, and others to understand ways that AI will influence everyday processes and outcomes. Many questions arise from prospective scenarios, such as what will happen when an AI outperforms humans on common tasks at work, and researchers are attempting to answer these questions regarding technical, practical, philosophical, and other aspects of AI. However, there exist few replicable examples—where an AI outperforms humans—beyond strategy games like AlphaGo, chess, the recent Cisero [7] that simulates diplomacy, and similar contests. We propose a conceptual model of an AI collaborative system comprising an HAI team similar to how such teams will operate. In addition, we present a replicable method and empirical evidence of HAI teams outperforming human–human (HH) teams on a concept development task. The study's findings reveal that participants who collaborated with an AI "teammate" and followed its suggestions generated a higher quantity of ideas. Moreover, these ideas were more frequently rated as "creative" by experts, compared to those produced by teams of two humans working on identical tasks.

In our model, we present a novel approach to developing a class of AI systems known as the Disruptive Interjector (DI). These systems observe human activities and interject with suggestions to stimulate idea generation or problem-solving within HAI teams. Unlike traditional creativity support systems, the DI replaces one human teammate in an HH team with an AI to form an HAI team. This approach encourages exploration of diverse possibilities rather than converging on a single correct solution, setting the DI apart from tutors, chatbots, recommenders, and other similar systems. We outline a conceptual design for the DI system and demonstrate its application using an open source system (https://experiments.withgoogle.com/autodraw) that is based on deep convolution neural networks (CNNs) [1].

By showing that HAI teams can outperform HH teams in design tasks, we provide compelling empirical evidence of AI's potential in collaborative tasks. As we advance into the next era of human-machine collaboration, this research contributes to the conceptualization of a new class of AI systems. These systems observe human actions and provide real-time suggestions, empowering researchers to design and test innovative concepts involving AI. Our demonstrated methods can be applied to explore various environmental contexts, such as distributed teams or mixed reality, and different modes of interaction during collaboration, including gestures and nonverbal cues.

*Corresponding author: Joseph M. Makokha, Department of Informatics, University of Oslo, Norway. Email: josepmak@uio.no

Figure 1
AI collaborator model



2. Background

Collaboration [8] is necessary for achieving high performance, and given today's work environments with a growing presence of AI agents, we anticipate humans to increasingly work with machines in HAI teams [6]. During such collaboration, exchanging ideas is a critical element and can be achieved using audio or visual modes. Given the recent advances [9] in computational tools and methods in AI and machine learning (ML), we propose an AI system that observes what a human is doing and then provides images, text, or voice suggestions based on their actions, as illustrated in Figure 1. In this paper, we implement such a collaborative AI that provides visual (image) suggestions in the moment, based on what a user is drawing. A variation of this setup can provide voice/text suggestions instead of images.

2.1. Design and operation of the Disruptive Interjector (DI)

The collaborative AI used in this study observes a person and interjects with suggestions without waiting to take turns the way humans do with each other, hence the "Disruptive Interjector" name. Conceptually, it comprises three parts: a human, an AI system trained to match images and which has access to an image database, and an interface that mediates the interactions between the two. The human in this case is involved in a cognitive task in which ideas on how to solve a specific task are generated and presented as sketches. A closer look at the other two components—the interface and the AI—suffices.

Interface: an interface such as a web-connected tablet is suitable for this kind of task, given that the generated ideas are sketched and captured sequentially. Capturing can be achieved in several ways such as having a remote monitor record entire sessions, transitions at the end of each idea just before the canvas is cleared for a new idea, or by having the user capture screens of each idea before resetting the canvas. Most important though is the ability of having an AI remotely receive sketch marks as the user proceeds with each sketch and then evaluate and respond with visual suggestions. In the current implementation, the suggestions are displayed on a designated area of the same interface, but separate interfaces can be used for sketching and displaying suggestions. In other alternative implementations where the AI uses voice to make suggestions, a speaker or a combination of speaker and text display may be used.

The AI: For this study, we used an AI fine-tuned using neural networks on a set of approximately 1500 sketches from human artists. The images spanned various categories of common items at home, work, play, and travel, as well as animals, plants, and many other groups. This AI matches an image with others in a database and then returns multiple images as suggestions based on a set of criteria. The criteria can be percentage accuracy of, say, 60% or greater, or this can be the top 16 matches, for example. The AI used for this study displays a varying number of suggestions from one sketch to another, so a configuration based on percentage accuracy suffices. It does not understand the task at hand nor the context, so it makes suggestions based on close matches of the sketch at any point with others in a remote database. This results in the AI sometimes suggesting wild, off-topic images/ideas that seem to be effective in inspiring new possibilities. While beyond the scope of this paper, our initial direction was to develop an AI tool to help engineers share ideas quickly and effectively, but we realized that they were using the tool for inspiration rather than adopting the suggested images for quick completion of their

ideas, hence repurposed the tool for early concept development in engineering design. This feature of sometimes suggesting wild, off-topic ideas points to a possible reason the AI outperforms another human collaborator.

2.2. Augmenting human abilities in technology-mediated tasks

The study of technology-mediated tasks highlights how digital tools have long supported creativity and collaboration, from early creativity support systems [10] to modern frameworks [11] for HAI co-creation. Prior research by scholars like Kempkes et al. [12], Tang et al. [13], Sinnemann and Weiss [14], and Abi Saad and Agogué [15] comparing computer-mediated communication (CMC) with face-to-face (FTF) collaboration shows that CMC tends to enhance idea generation (divergence), while FTF more effectively filters and refines ideas (convergence). Concepts such as representational gaps and "pivot thinking" underscore the need to balance divergent and convergent processes, a role that emerging AI systems may uniquely fulfill by dynamically bridging these gaps. Building on earlier tools like the Problem Formulator [16] and Baya et al.'s [17] Electronic Design Notebook, recent systems such as Yun et al. [18] demonstrate how technology can augment ideation. Our proposed DI extends these approaches by using AI to spontaneously generate prompts like images, phrases, or suggestions that emulate human collaboration and stimulate new solution pathways in real time.

In fields such as engineering and business, teamwork is essential for tasks like concept design and ideation. These activities, typically performed by groups of two or more people working simultaneously, often lead to innovative ideas and products. Given the value of creative ideas across various domains, researchers from disciplines such as engineering [19], psychology [20], art [21], and others have investigated the conditions and environments that foster creativity. Creativity is often defined as the ability to generate new, surprising, and valuable ideas [21] or to find nonobvious solutions through reframing problems [19]. Their findings highlight the potential of technologies like AI to enhance human cognitive abilities in tasks like concept development within engineering contexts. This concept has long been anticipated by scholars and researchers. For instance, Licklider [22] predicted that future human-machine partnerships would excel at intellectual operations, while Simon [23, 24] foresaw machines capable of performing a variety of thinking tasks similar to those done by humans.

With regard to computers and AI, there are many perspectives that can guide our approach to designing, developing, and deploying these systems in real-world applications. Sheridan and Verplank [25], as well as recent ideas by Richardson et al. [26] and Moruzzi and Margarido [27], for example, provide a model for shared control between humans and machines in the context of undersea vehicles that presents one way of visualizing HAI collaboration.

The AI can extend, relieve, back up, or replace humans depending on the circumstances, and we see this applied, for example, in conventional driver assistance systems, though different in context. Figure 2 shows an outline of HAI engagement based on their model, and we build on the "Extending capabilities" aspect in this study.

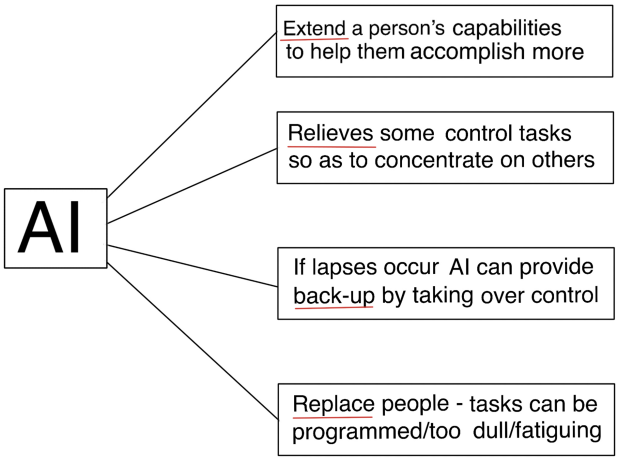
2.3. Measuring effectiveness of AI in engineering teams' concept design

To make any inferences on the effect of a technology intervention on team outcomes, we require appropriate methods of measuring these outcomes—for example, engineering design concepts expressed through sketches. As might be expected, there is a rich body of literature concerning this phenomenon. Arnold [19], Dow [20], and others [28–32] have explored conditions and environments that enhance people's creativity—defined, for example, as their ability to come up with new, surprising, and valuable ideas [21, 33] or finding solutions that are far from obvious through a shift in problem formulation [34]. Despite the disagreements on the definition of creativity [35], there has been convergence in attempts to define it, straddling the following three aspects:

1) The Gestalt position based on insight and productive thinking that arises when one grasps the essential features and their relationship to the solution, as proposed by Wertheimer in [36]
2) The person as emphasized by psychologist Guilford [37]
3) Characteristics of a product such as producing "effective surprise" combined with a "shock of recognition" as viewed by Bruner [38]

Others like Shah et al. [39] and Fiorineschi and Rotini [40] associate creative design with idea generation and developed a criterion for novelty, variety, quality, and quantity of ideas. Furthermore, they have introduced methodologies for systematically classifying design sketches generated in response to specific prompts, emphasizing solution-focused ideation that evaluates both the output and its functional attributes. A different take previously advanced by Weisberg and presented in Pichot et al. [41] proposes that "a product is creative if it is novel, and if it was produced intentionally," regardless of whether it is of value to anyone. He credits Csikszentmihalyi's influential analysis of creativity as having catapulted value into contemporary definitions [42] of creativity. However, a different suggestion from Simonton [43] takes the definition from the United States Patents and Trademarks Office (USPTO), which uses three criteria: new, useful, and surprising (nonobvious), as a definition of creativity. From this brief review of creativity definitions, it is clear that the chosen definition will vary depending on the domain, whether it be engineering, product design, psychology, or art. Cascini et al. [44] add to Gero's [45] previous work by distinguishing between creativity residing in the artifact itself, which is then evaluated by society, and the idea that certain processes can produce artifacts deemed creative. Relevant to our discussion is the assertion that while the "creativeness" of an artifact is judged by society,

Figure 2
A model for human–AI collaboration



understanding these processes could help us learn how to foster creativity in individuals.

In this discussion, creativity will be considered from a product lens by evaluating the outcome of novice engineering designers' activity when asked to generate and sketch ideas on how to accomplish given tasks. We narrow down to three characteristics comprising aspects from the above discussion on past efforts to describe creativity:

1) Weisberg's [46] proposal that a product is creative if it is novel and if it was produced intentionally, regardless of whether it is of value [47] to anyone
2) One that exhibits "uncommonness" compared to others in the group, in addition to being adaptive to reality as characterized by Barron [48] or
3) As put differently by Newell et al. [49], one that demonstrates novelty and appropriateness or value.

For this study, participants engaged in a controlled experiment to create tasks that specify both an intended output (e.g., "a way to get a toddler into bed") and a corresponding set of functional requirements (e.g., "have a baby inside a bed"), which are then systematically evaluated. Expert ratings is another viable approach, though Shah et al.'s [39] systematic approach with its methodical, consistent procedures has the benefits of producing more comparable and replicable results [50].

2.4. The Disruptive Interjector (DI)—a different kind of AI system

We examine a class of AI systems, illustrated in Figure 1, which monitor human actions—such as sketching or conversation—and periodically interject with visual, textual, or vocal suggestions intended to guide the designer toward novel and promising directions. This represents one of several potential modes through which AI can enhance human capabilities in ideation and concept development. To distinguish this class of computational artifacts from others operating at a similar level of abstraction such as recommender systems or cognitive tutors, we introduce the term "Disruptive Interjector (DI)."

In other words, this is an interactive "hint giver or AI Collaborator" that is distinct from "chatbots/voice assistants," hence the name DI that characterizes the unsolicited interaction initiated by the system during collaboration with a human. Several related but distinct concepts operate at a similar conceptual level as the DI system. These include cognitive tutors, recommender systems, priming and anchoring, planning systems, and design systems, as summarized in Tables 1 and 2 [51].

The following distinctions clarify how these systems differ from the DI:

1) Cognitive tutor: operates with a higher level of expertise than the learner, guiding them toward correct procedures. In contrast, the generative AI tool functions as a design partner. Its knowledge may be equivalent to or even lower than that of the human collaborator.
2) Recommender system: relies on a user's prior selections to infer preferences and generate new suggestions. While both systems depend on user interaction, the generative AI tool exhibits limited contextual understanding, instead predicting and evolving sketches in real time to suggest novel or tangential ideas that may inspire new directions.
3) Priming/anchoring, planning, and design systems: depend on independent sources of input rather than a user's own

Table 1
Differences between AI generative partner (Disruptive Interjector) and other systems that observe/track user input and then offer suggestions

Tool	Goal	Method	Direction
Disruptive Interjector (DI)	Lead a novice designer to imagine new ideas	Tracks a user's actions and interjects with suggestions	Divergence
Cognitive tutor	Guide the learner based on their performance or level of understanding	Observes user actions and adapts subsequent content to align with defined learning objectives (applicable to routine procedural tasks)	Diverge/converge
Recommender system	Offer the user one or more options from a larger collection, based on their previous choices	Utilizes user preferences to generate targeted recommendations that satisfy specified criteria (for simple selection tasks like choosing a movie to watch)	Convergence

Table 2
Differences between AI generative partner (Disruptive Interjector) and other systems that offer suggestions without needing a user's input

Tool	Goal	Method	Direction
Priming/Anchoring	Strategically offer a starting point to influence the outcome toward the desired one	Influence a person's decisions using a given starting point of information	Convergence
Planning system	Narrowing down to a solution offering the best outcomes from a complex set of options using available information	Determine combinations resulting in the outcome offering the greatest benefits using system metrics	Convergence
Design system	Given a large set of options, optimally identify solution(s) based on a preferred criterion	Reduce to the lowest cost difference between the starting and final points	Convergence

selections. This contrasts with the DI, cognitive tutor, and recommender system, which adapt dynamically to the user's actions and evolving design process.

Notably, the generative AI system aims to diverge, as opposed to the others, which converge toward a solution or answer. This kind of divergence is desirable in the course of idea generation, given that it may help resolve the "idea-fixation" [52–55] phenomenon that inhibits people from considering other possibilities.

3. Methodology

The methodology here provides insights into design considerations for creating a collaborative AI system that functions as a partner in an HAI team, especially as AI becomes increasingly integrated into various domains. It then proceeds to offer a replicable method for conducting studies involving AI that yield empirical evidence demonstrating that AI can outperform humans as collaborative partners in cognitive tasks like early-stage concept development.

3.1. System design and implementation of the Disruptive Interjector (DI)

This section aims to provide one implementation of the AI system introduced earlier in Figure 1 and note that other implementations involving a voice/text conversational AI were explored (though beyond the scope of this paper, so not discussed here). The system comprises an interactive interface such as a touchscreen, on which a human makes sketches that are relayed in real time to a computer program—an AI in this case. The AI matches the sketch with others in a curated library and then returns and displays a subset of suggestions from the collection.

3.1.1. Training dataset for the AI

The training dataset is made up of 1540 images created by professional artists involved with the Google AI experimentAutoDraw. While these sketches are relatively artistic with some level of detail, they nevertheless work well for training the AI to detect the in-the-moment human sketches created by participants in our study, where ideas are generated with a focus on quick, low-fidelity "sketchy" images as opposed to detailed drawings. However, this is a relatively small number of sketches compared to contemporary pre-trained classifiers. Figure 3 shows a sample from the collection used to train the AI in AutoDraw.

3.1.2. Training algorithms incorporating pre-trained ResNet-50

The ResNet-50 is a pre-trained deep CNN [56] used in image recognition applications and has been applied—through transfer learning—to many image classification tasks such as fault diagnosis in manufacturing [57] and in the classification of malaria cells [58]. In a simplified form for our use, we take the first 49 layers of the ResNet-50 and add two new layers—a fully connected layer and a Softmax classifier that are used to fit the class labels of the provided sketches.

The next step involves training a neural network to map new images to the closest matching image(s) in the dataset. Figure 4 outlines the process, which proceeds iteratively between data preparation, training, and evaluation stages. Once a sufficient level of accuracy is obtained, the model predicts which of the sketches in the existing database matches one that a human is sketching in real time.

Typical hyperparameters for this ResNet-50-based system are AdamW optimizer; learning rate of 3e-4 (cosine decay); batch size of 32 or 64; 20, 40, or 60 epochs; and CrossEntropyLoss as the loss function.

Figure 3
A sample of sketches used in training the AutoDraw AI

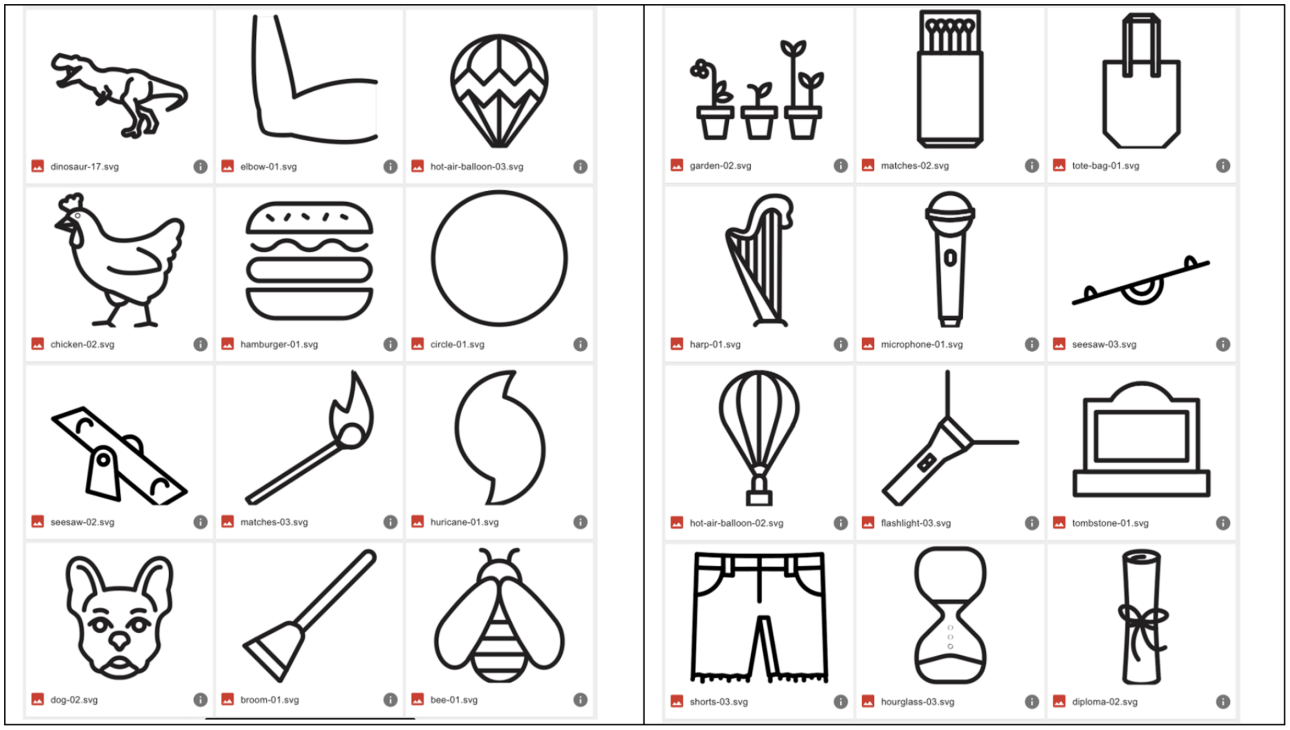
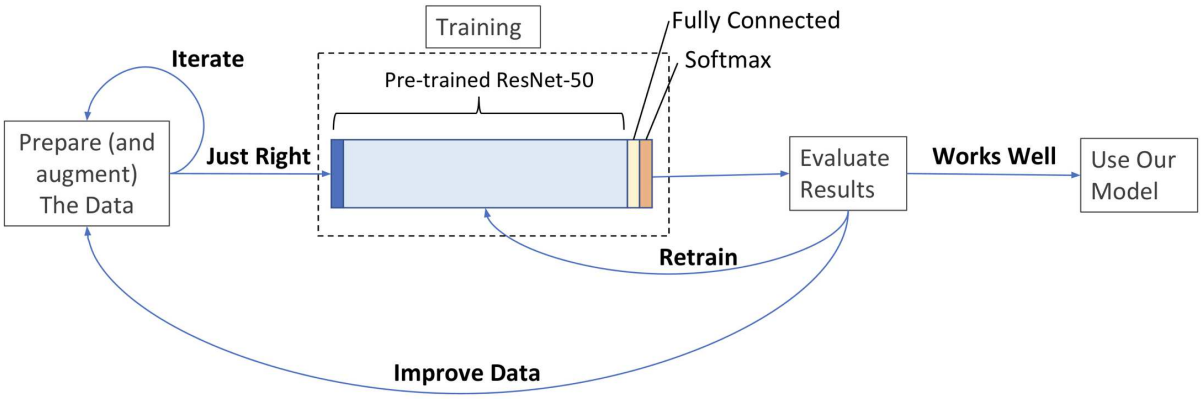


Figure 4
Outline of process for training the AI



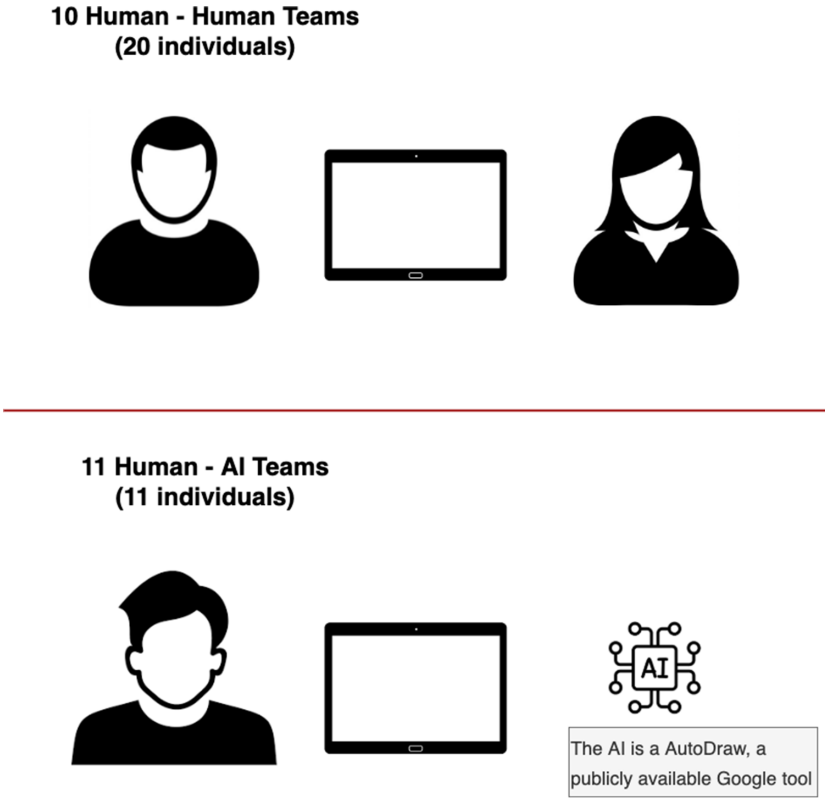
3.1.3. Customization

In most image classification tasks, a single correct answer is often desired for decision-making—such as recognizing a face to grant access and a street sign or other object to determine course of action, among others. For our system design of the AI, we seek multiple suggestions from the database as opposed to the one best prediction. This presents several options: we can restrict the output to, say, the top 10 or take all candidates with an accuracy above 50%, for instance. The latter is preferable given the task at hand, since we are using the system as an inspiration for new ideas, as opposed to finding a correct answer.

3.2. Study setup

Thirty-five senior undergraduates and beginning graduate students participated in the study exploring the impact of an AI as a collaborator in place of another human. They were randomly placed in either an HH team or an HAI team, as illustrated in Figure 5. Within these participants, 27 (77%) were men, and 8 (23%) were female, aged 21–32 years, with a median of 23. They were enrolled in cross-disciplinary programs including engineering and design, among others, and were familiar with using sketching to express ideas in their classes. The level of engineering design experience and sketching to express ideas was thus controlled for.

Figure 5
Team configuration for human–human (HH) and human–AI (HAI) teams



They had a wide variety of national and cultural backgrounds including the USA and were fluent in English. They participated in accordance with the University's Institutional Review Board (IRB) approved protocol.

Participants received orientation and then completed two consecutive design tasks, each lasting approximately 7 min. In the HH condition, the AI sketching assistant was turned off. Each session began with an initial brainstorming phase, after which participants selected a final concept from their generated ideas.

The design prompts were crowdsourced from a diverse pool of novice and experienced engineering designers who contributed examples of meaningful but challenging real-world problems they had recently faced. From more than two dozen submissions, two prompts were chosen for their abstractness, general relevance, and potential for reinterpretation:

1) Design a way to get a toddler into bed.
2) Design a way to transport your computer on campus.

Following the sketching activities, participants completed a questionnaire and an exit interview aimed at understanding their design reasoning, collaborative dynamics, and interaction experiences with either a human or an AI partner. Additional data collected encompassed participant demographics (age, design training, cultural background, and teamwork preferences), timing data for each session (start and end times), sketch counts and timestamps, and the type of collaborator (human or AI). Finally, two trained evaluators independently assessed the resulting sketches for novelty and usefulness [27].

3.3. Study procedure

Participants were invited to take part in the study via email. The experiment was carried out in a dedicated room on a university campus designed for human-subject research, as well as in an off-campus facility approved by the university's IRB. Prior to participation, each individual read and signed a printed consent form. They then received an orientation on how to use the web-based sketching application on a tablet. The procedure differed slightly depending on whether participants were paired with another human or an AI agent. All participants practiced using the sketching tool and learned how to capture screenshots of their ideas before beginning the main study.

After orientation, participants completed two consecutive 7-min design tasks. The first required designing a solution for another person, followed immediately by the second one that focused on designing for themselves. Each session began with idea generation and sketching, followed by the selection of one final concept representing their best idea. Participants used a stylus for sketching, with one stylus per person in HH pairs and one shared stylus in HAI pairs. The sketching web application ran on the same platform for both conditions, with the distinction that the AI suggestion feature was disabled for HH teams. Table 3 summarizes the specific instructions provided to participants.

Participants were notified midway through each task and again 1 min before the end to help them manage their time. The second task began immediately after the first. A few participants (approximately three) asked whether they could include "any" ideas—specifically, unconventional or "crazy" ones—and were informed that such ideas were acceptable. While this clarification may have slightly influenced their creative exploration, the impact was considered minimal, as these ideas still needed to demonstrate usefulness, consistent with the study's operational definition of creativity.

Afterward, they filled out a questionnaire capturing demographic information, level of design experience, and preference for collaborative work. This was followed by a semi-structured team interview designed to explore the strategies used in idea generation and selection and to compare their experiences collaborating with either another human or an AI partner. Participants were asked two central questions:

1) To describe the process, they followed when generating ideas in collaboration with their partner (or the AI agent, for HAI teams).
2) To explain how they selected their best idea and the strategies that guided each of these steps.

3.4. The data collected

Throughout the study, we recorded the start and end times of each design round, the number of sketches produced along with their timestamps, and whether the participant collaborated with a human partner or an AI agent. After completing the sketching sessions, each participant filled out an individual questionnaire in a private setting. This questionnaire gathered information on demographic attributes such as age, level of design training, cultural background, and self-reported affinity for teamwork.

Following the questionnaire, participants took part in a semi-structured interview that explored the strategies they employed in

Table 3
Instructions given to HH and HAI study participants

Human–human (HH) pair	Human–artificial intelligence (HAI) pair
1. You will collaborate with a partner on a design task	1. You will collaborate with an intelligent computer app on a design task
2. You have 2 tasks to complete, with 7 min for the first task and 7 min for the second	2. You have 2 tasks to complete, with 7 min for the first task and 7 min for the second
3. Draw/sketch as many options as you can come up with, saving each sketch before starting the next version. Both quantity and quality of ideas matter. Save each drawing by taking a screenshot on the tablet before starting the next one	3. Draw/sketch as many options as you can come up with, saving each sketch before starting the next version. Both quantity and quality of ideas matter. Save each drawing by taking a screenshot on the tablet before starting the next one
4. You may accept each other's ideas as is, decline them, or combine them with yours to make new ones as you see fit	4. You may accept the app's suggestions as is, decline them, or combine them with your own to make new ones as you see fit
5. Let's practice using the app	5. Let's practice using the app

Table 4
Summary of data collection

Stage	Activities
Recruitment and Invitation	Participants were invited via email containing study details, location, and researcher contact information
Consent and Onboarding	Participants reviewed and signed informed consent forms and were briefed on the study procedures
Tool Orientation and Practice	Participants received an orientation to the tablet-based sketching application, practiced sketching and capturing screenshots, and learned task procedures; orientation varied slightly by collaboration condition (HH vs HAI)
Design Task 1 (Design for Others)	Participants completed a 7-min design task involving idea generation through sketching and selection of a final concept; time reminders were provided
Design Task 2 (Design for Self)	Participants immediately completed a second 7-min design task with a new prompt, following the same sketching and selection process
Questionnaire	Participants completed an individual survey collecting demographic information, design experience, cultural background, and collaboration preferences
Semi-Structured Interview	Teams participated in a short interview exploring idea generation strategies, idea selection processes, and collaboration experiences with a human or AI partner
System and Artifact Logging	Automated logs recorded task timing, sketch timestamps, collaboration condition, and final idea selection duration
Data Archiving and Anonymization	All sketches, logs, survey responses, and interview recordings were anonymized and securely stored with date and time stamps

both idea generation and idea selection across the two tasks. The final dataset comprised:

1) 263 sketches (ranging from 8 to 15 per team),
2) Time logs showing how long participants took to choose their final idea,
3) Individual survey responses (one per participant), and
4) Audio recordings of debriefing interviews (one per team, lasting 1–5 min).

All data were anonymized and stored with date and time stamps. Table 4 summarizes the stages and activities involved in the data collection process. This is the data that was eventually rated for creativity based on two metrics—usefulness and novelty.

4. Experiment Results

The complete set of sketches underwent an initial data validation process to identify and correct any errors or anomalies. This review revealed instances of invalid data from teams that did not adhere to the study guidelines, for example, by sketching multiple ideas on a single canvas. Additional irregularities included duplicate screenshots of the same sketch and test marks created while experimenting with features of the AI sketching application. All such sketches were excluded from subsequent ranking and analysis to ensure the integrity of the dataset.

Figure 6 presents a random sample of sketches for the two tasks, taken from a collection of 263 sketches, while Figure 7 shows a sample from HH teams beside those from HAI teams. When these sketches are rendered in black and white (as was the case in printouts given to the raters), it is difficult to identify which of the 21 teams sketched any of the ideas, limiting bias on perceived team composition. For the next step, these 263 annotated sketches (printed on a 3 × 5 index card in black and white) were given to the trained research assistants for rating. The annotations briefly stated what the idea in the sketch was about—based on exit interview notes, for example, "an arm that picks up baby." Using experts or trained judges is a common method for evaluating sketches that represent design ideas [59, 60].

4.1. Training the research assistants (judges)

The research assistants received training on how to evaluate the sketches to ensure that they were looking for the same characteristics and assigning scores consistently. They were shown some examples and discussed ways they would group and rate them. Among other aspects covered in the training were:

1) Getting familiar with the entire collection of sketches
2) Agreement in sorting the sketches into groups of similar ideas
3) Ability to consistently assign scores of usefulness and novelty, the latter based on uncommonness from the other ideas in the collection
4) Ways to resolve the grouping or rating of ambiguous ideas
5) How to identify ideas that met the "usefulness" measure
6) Ways to reach a consensus when large disparities occur in grouping or rating of ideas

4.2. Evaluating sketches: Step 1

The first step involved looking at all the ideas and sorting them into groups of similar ideas. This revealed disproportionate groups of ideas (e.g., steps/ramps for the first task and bags for the second one), which comprised about half of the sketches for each task. These were deemed very common (not novel). Sketches from two HAI teams could not be evaluated the same way as the others, as they had sketched complex ideas by combining more than one idea on the same screen. These complex sketches were therefore set aside. Next, a small number that were termed "nonsensical" by the research assistants due to their inability to solve the given task (hence not useful at all) were also set aside. One more group was set aside as it was discovered that they knew each other, collaborated frequently on problem-solving tasks, and had

Figure 6
Sample sketches for the 2 tasks—(a) ways of getting a toddler into bed and (b) transporting your computer on campus

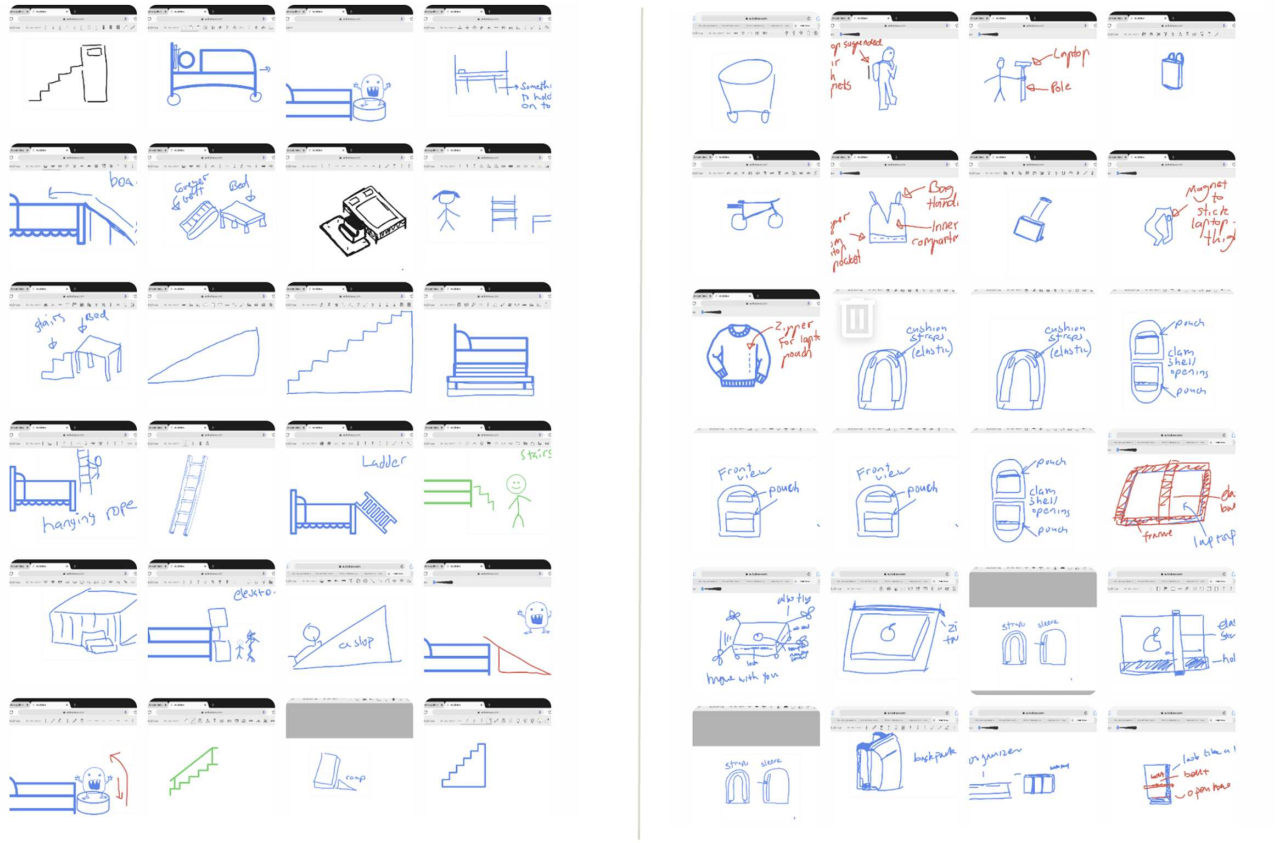


Figure 7
Sample sketches from human–human (HH) and human–AI (HAI) teams

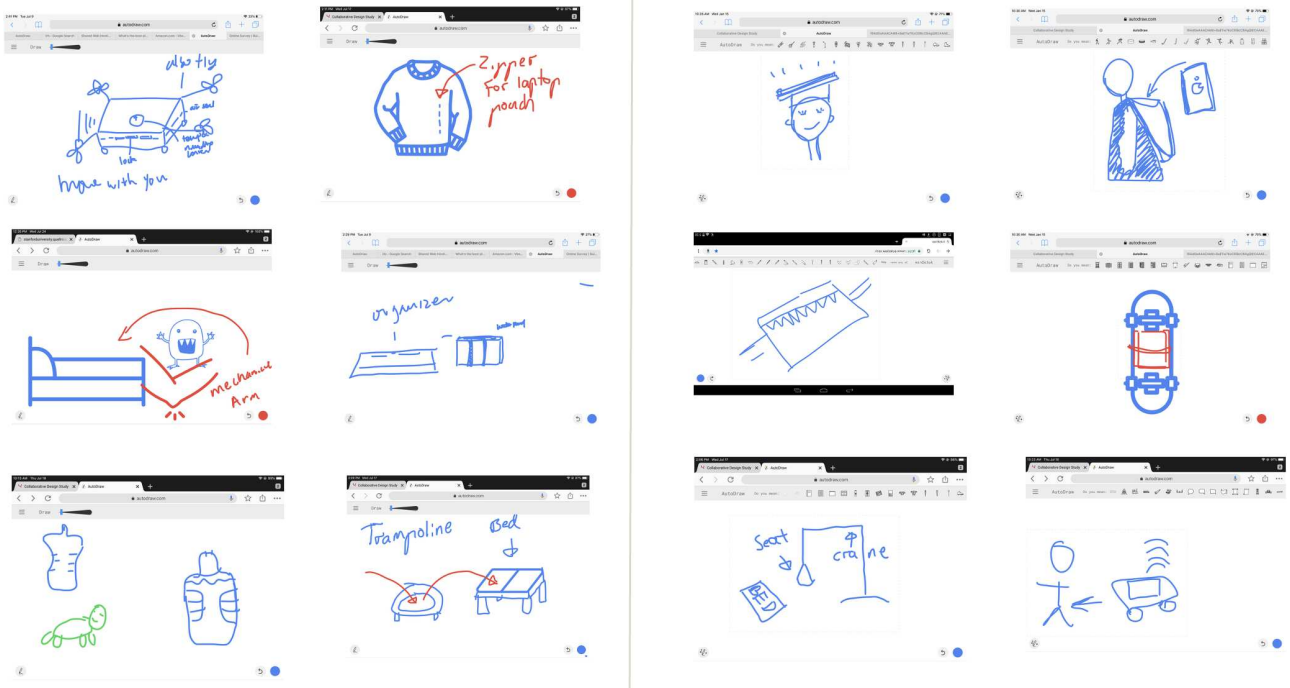


Figure 8
Rating scheme used on the sketches

expert-level design experience—making them quite different from the rest of the participants. This left us with sketches of ideas that demonstrated both usefulness and novelty, an important requirement for our chosen measure of creativity. The sketches fell into the following groups that were created by the raters:

1) For the toddler task, we had enticing with food, enticing with other things, scaring them, propelling, ladders, suspending the baby in the air, bed that moves, picking up the baby, and others not in any of these groups.
2) For the computer task, we had pets, folding kind, suspended in air, wearables, ground vehicle, gravity or pressure, and other individual ideas not in any of these categories.

These 16 groups (9 for getting the toddler to bed and 7 for transporting a computer on campus) were used later to determine the variety of sketches in each team.

4.3. Evaluating sketches: Step 2

In the second step, the raters were instructed to rate two characteristics—usefulness and novelty—of every sketch on a 4-point Likert scale (note that a 5th point of zero was assigned to sketches identified as "not useful" as well as those grouped as "not novel"). The raters were therefore interested in how useful the idea was at solving the task and how novel (different from the other ideas in the task category) each idea was. For usefulness, 1 point was assigned to a slightly useful idea, 2 points for moderately useful, 3 for very useful, and 4 for extremely useful. Sketches that were grouped as "nonsensical" by the raters were assigned a zero score for usefulness. In terms of novelty, a 1 was assigned to sketches that were somewhat similar to others, a 2 to those neither similar nor different from others, a 3 to those somewhat different from others, and a 4 to very different from others. The sketches previously grouped under "common" were assigned a score of zero for novelty.

Figure 8 shows the scheme that was applied. The interrater reliabilities for the combined sketches deemed both useful and novel (on ways of "getting a toddler to bed" and "transporting your computer") were computed. We used two evaluation methods—Krippendorff's and interclass correlation (ICC)—and obtained 0.68 for novelty taken alone, while the combination of usefulness and novelty gave a 0.60 reliability rate. The ICC options were model = "twoway," type = "agreement," unit = "single," and percentage confidence = 95%.

Creativity of the ideas was therefore based on the two characteristics, namely, "usefulness" and "novelty," as rated by trained assistants familiar with the research domain. They applied a scale of 1–4, with 4 as the highest.

4.4. Experimental—validating the ranking scale using an AI classifier

The first validation of ranking involved ICC and Krippendorff's score of the two trained research assistants who rated the sketches based on whether the sketched idea could fulfill the task (usefulness/value), as well as its novelty (uncommonness).

The second validation approach employed an AI-based image classifier developed using Google's Teachable Machine. The model was trained on a subset of 36 sketches that had received identical rankings from all three human evaluators, ensuring a consistent training reference. It was then tested on 10 unseen sketches drawn from the same distribution as the training set. The classifier's predictions matched the human raters' evaluations with 100% accuracy, confirming strong alignment between automated and expert assessments. Figure 9 illustrates two representative examples of the classifier's output.

This is a promising strategy for quickly classifying sketches in settings like engineering design classrooms, where students produce multiple sketches/concepts every session, making it difficult for the instructor to see every sketch in the moment. Such a classifier could be used to highlight samples in the moment for the instructor. The point here is to highlight an AI solution to classifying the sketches in addition to the human option that we employed in this study.

4.5. Analysis

To understand the impact of collaborating with an AI agent instead of another human on engineering concept generation tasks, we evaluated the sketches collected from the teams on a range of metrics including:

1) The total number of sketches in each team
2) Average number of sketches per team
3) Average number of sketches for the HH and HAI groups
4) Ratings of each sketch on novelty over a 5-point scale
5) Number of sketches receiving a high score from both raters in each team
6) Variety based on how many of the identified groups (9 groups in the first task and 6 in the second) were represented in a team's ideas

The effect of collaborating with an AI rather than another human can be evaluated since the teams had the same length of time for the assigned tasks and used the same interface for sketching ideas—a tablet with the presence of an AI assistant in the background for HAI teams or a disabled AI for HH teams.

Each of the 21 teams produced from 6 sketches to 23 sketches, as shown in Figure 10. This represents an average of 10.3 sketches over all teams.

Other values are shown in Table 5.

Figure 11 shows a breakdown between HH and HAI teams, with the HH teams averaging 8.5 sketches while the HAI teams average 11.9 sketches. The means and modes for the HAI teams were about 50% higher than those for the HH teams.

4.6. Cumulative ratings of sketches for the HAI and HH teams

Figure 12 shows the count of how many times the ideas from the HAI teams received ratings of 1, 2, 3, or 4 from the 2 raters. An idea (or sketch) could receive the same score twice if both raters gave a similar score.

Figure 9
Sample output from an image classifier

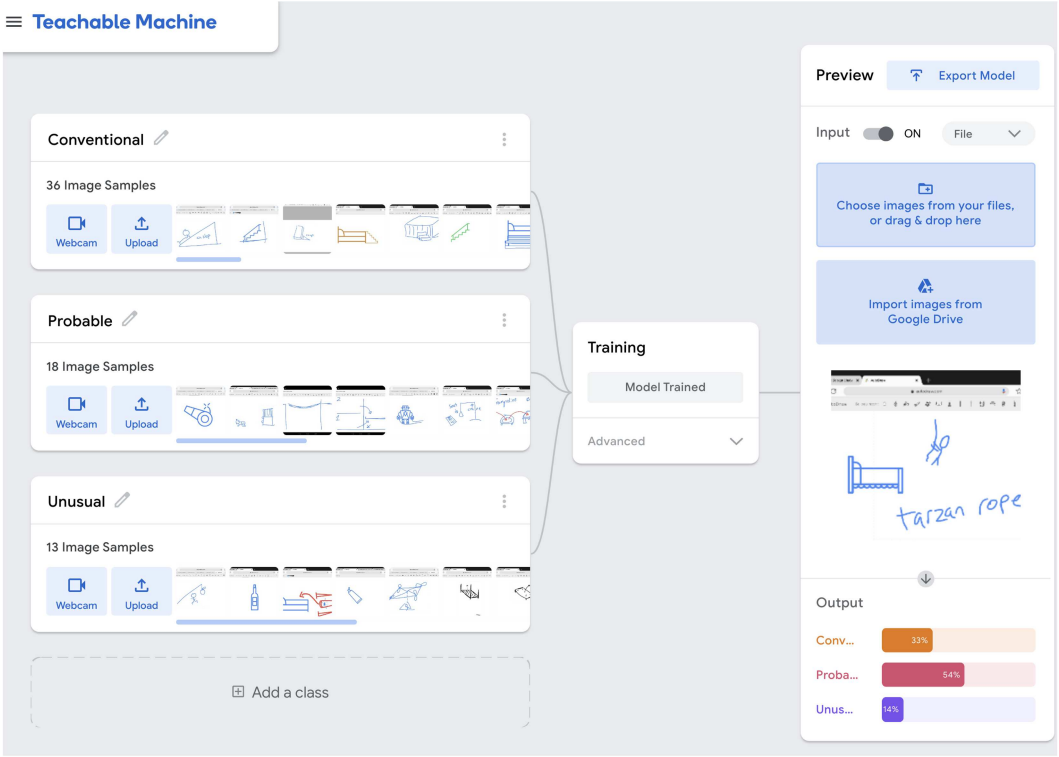


Figure 10
Proportion of sketches produced by the 21 teams

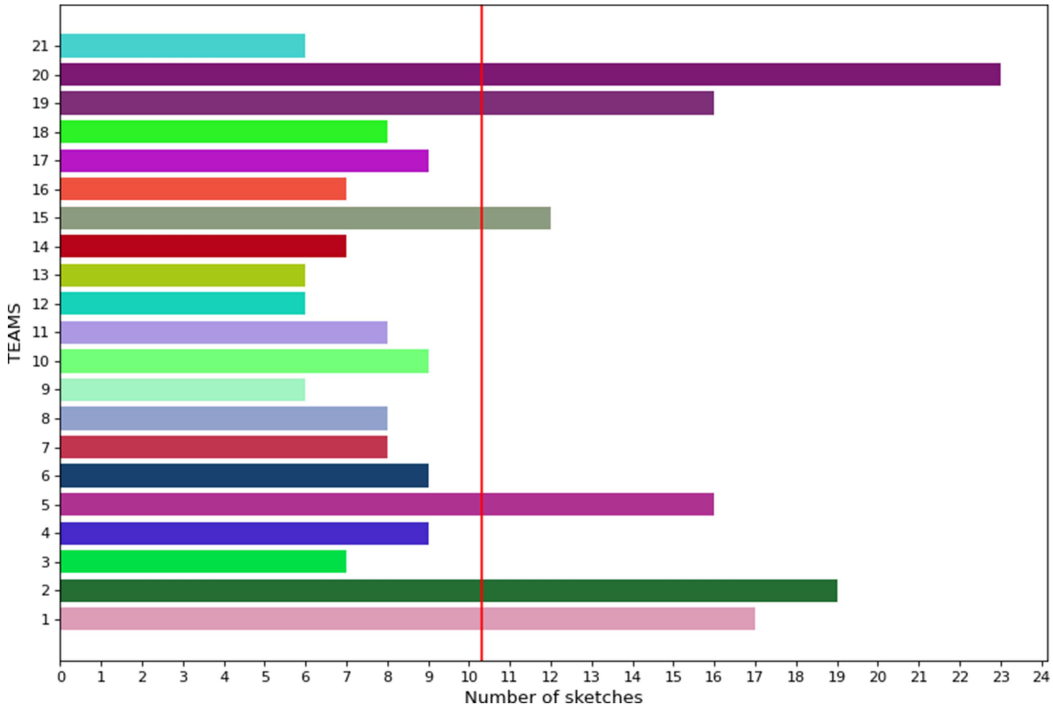
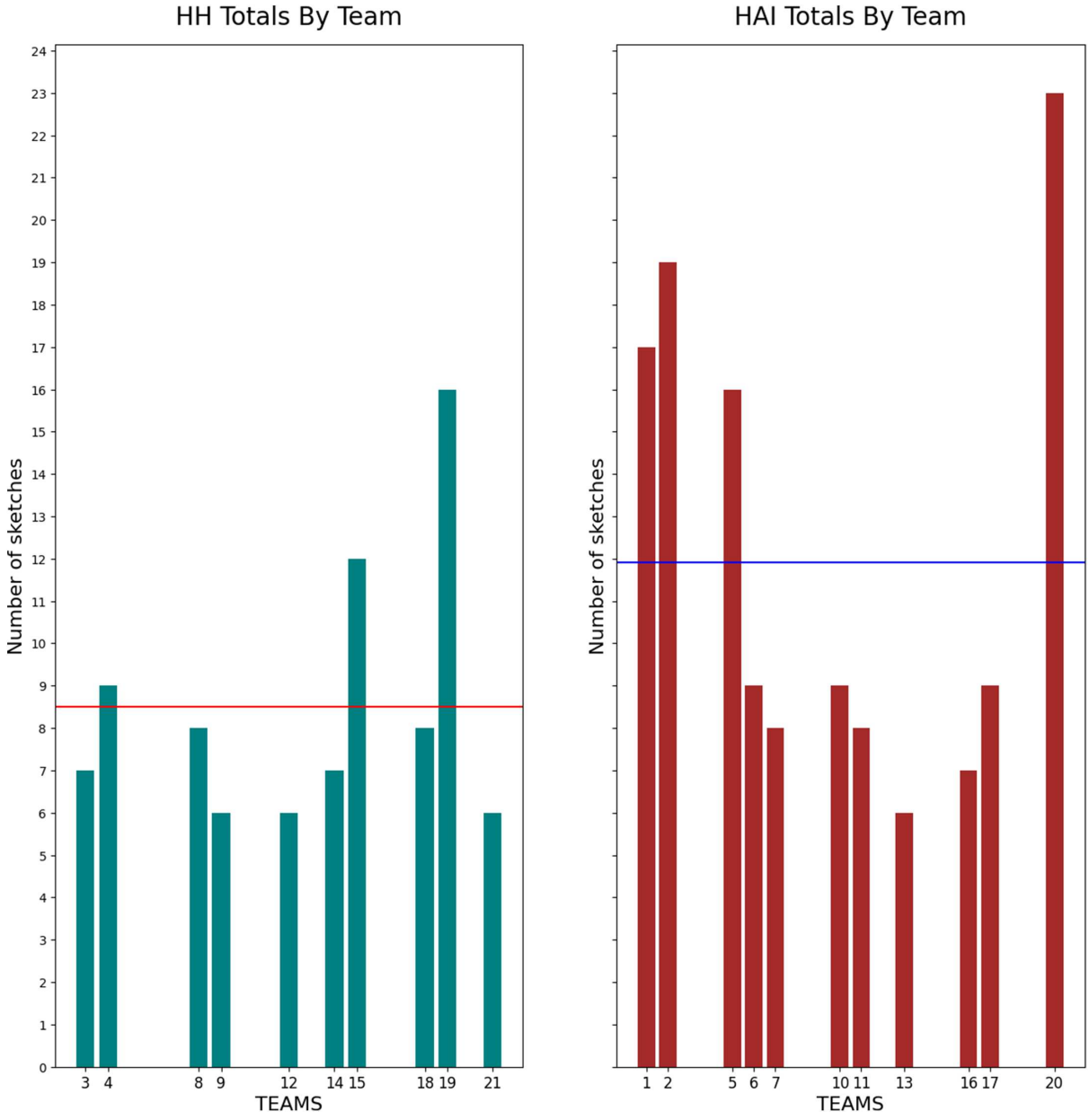


Table 5
Mean, mode, SD, and variance of sketches from the HH and HAI teams

	Mean	Mode	Standard deviation	Variance
HH teams	8.5	6	3.2	10.3
HAI teams	11.9	9	5.8	33.1

Figure 11
Total number of sketches by HH and HAI teams



Similar to the preceding section, a count of how many times the HH sketches received a rating from 1 to 4 is shown in Figure 13. Figure 14 shows the combined count of scores for the HH and HAI teams in one graph. The HAI teams have higher values in general, both the highest ratings and the total number of counts.

For the 122 sketches and 2 raters assigning values for novelty and usefulness, 488 distinct scores emerge. Figure 15 shows the total number of sketches from the HH and HAI teams that received a score of 1 (low), 2 (medium), 3 (high), or 4 (very high) from the two raters. It is evident that a larger number of the ideas (64%) received a 1 (lowest) or 2 compared with those receiving a 3 or 4 (highest), which take 36%.

The normalized sum of ratings attained by the HH and HAI groups is shown in Figure 15. This representation reveals that the HH teams had a larger proportion of the lower scores (1 and 2)

Figure 12
Count of ratings from 1 to 4 for the HAI teams

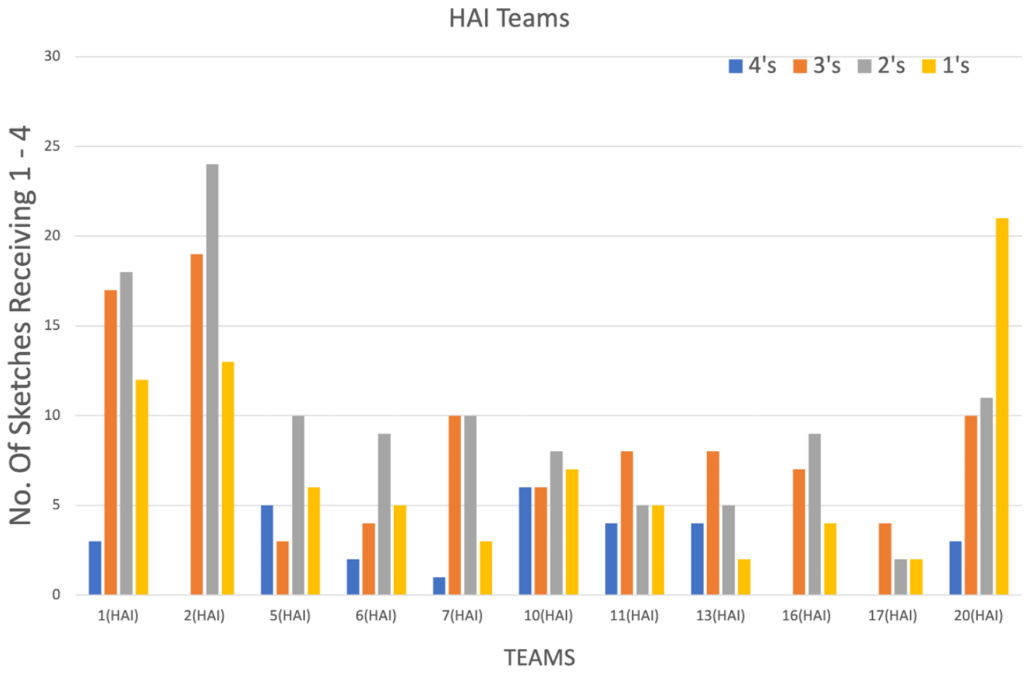


Figure 13
13 Count of ratings from 1 to 4 of the HH teams

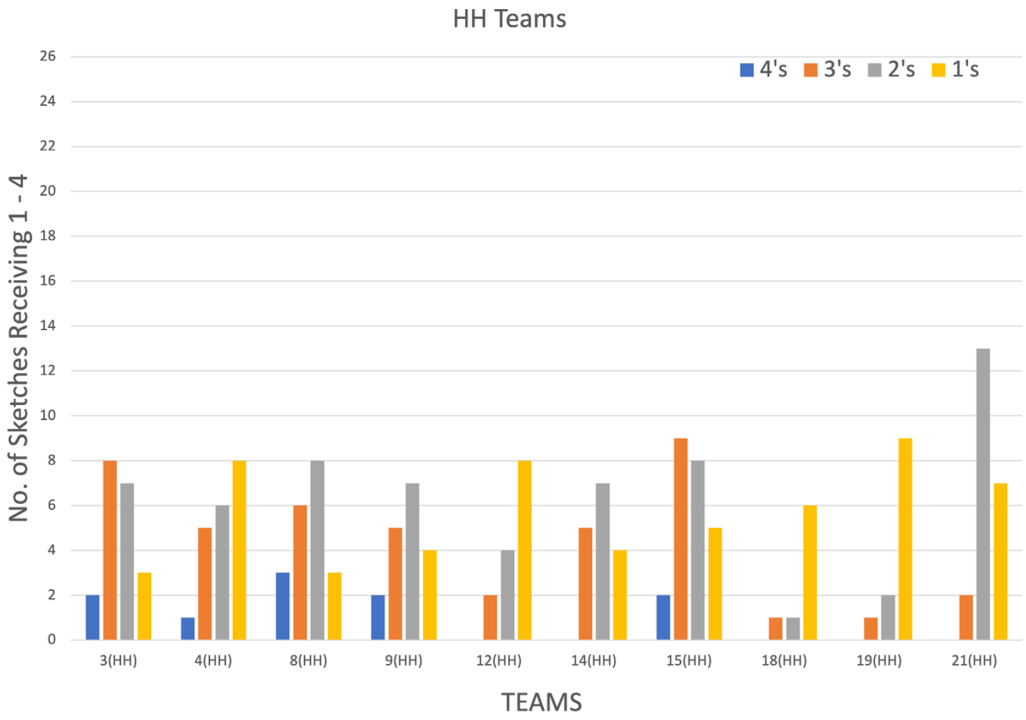


Figure 14
Count of ratings from 1 to 4 of the HH and HAI teams

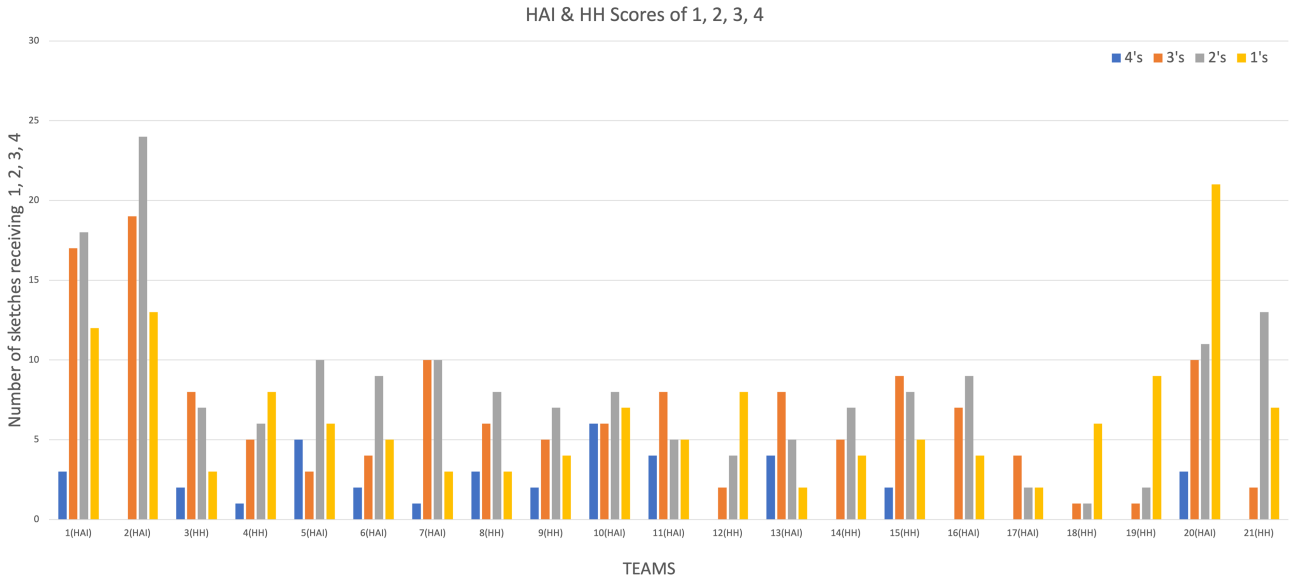
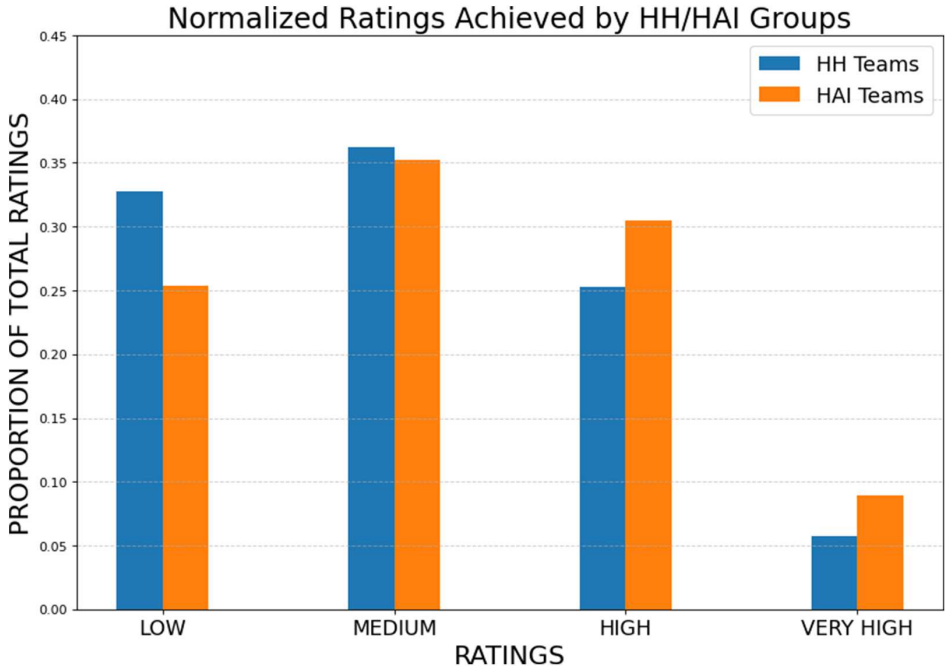


Figure 15
Normalized sum of ratings for the HH and HAI groups



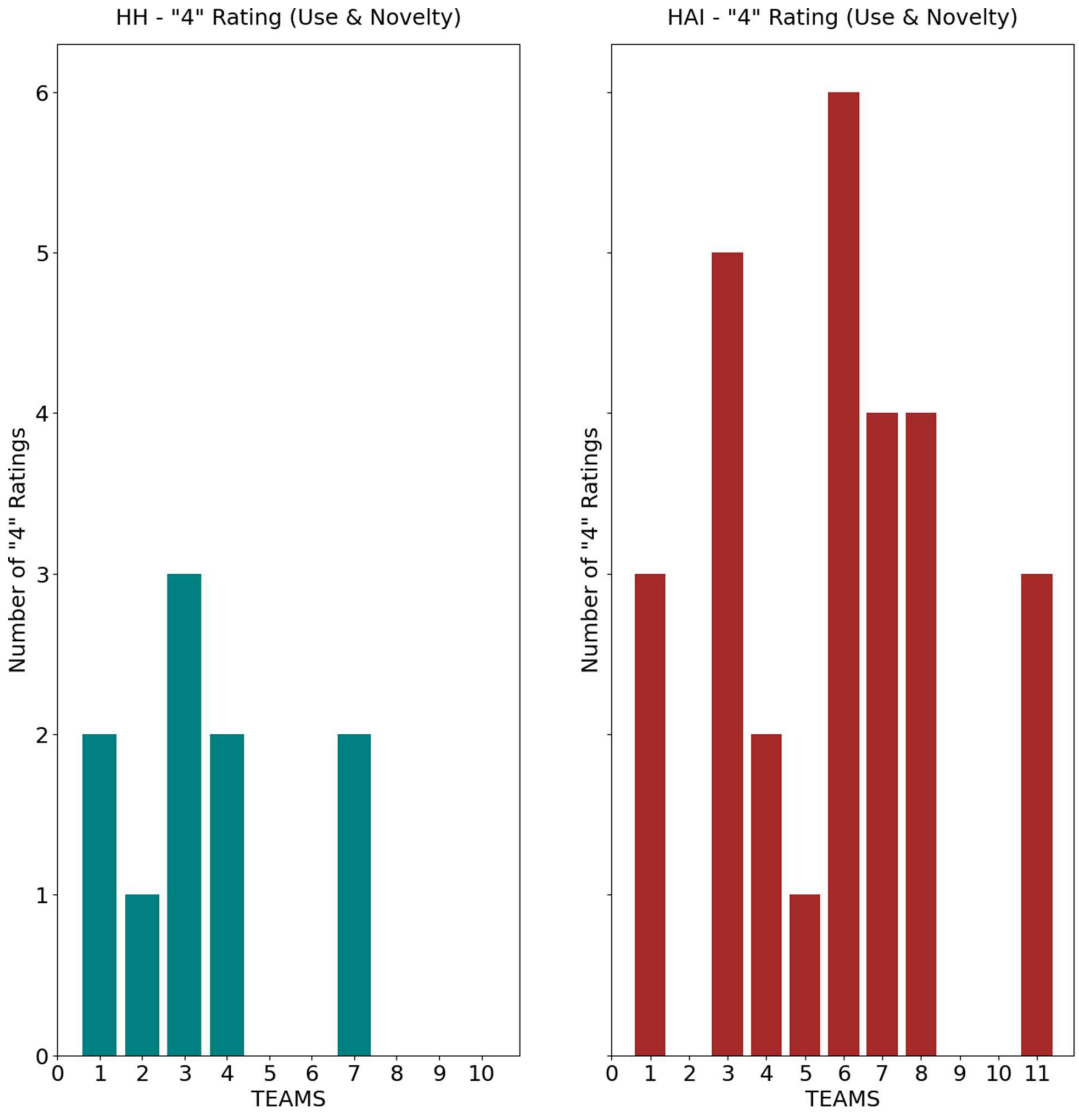
compared to the HAI teams that had more of the higher scored ideas, as might be interpreted from Figure 14 showing the count from all teams.

4.7. Focusing on the highest rated ideas

There are several ways to characterize the results presented above, including a consideration of every idea, lowest/highest rated ideas, most common/uncommon ideas, combining usefulness and novelty scores, etc. We chose a definition of creativity comprising usefulness, which is the ability to perform the specific task assigned, and novelty relative to other ideas in the collection. We therefore focus on the highly rated ideas as these are more likely to be candidates for creative ideas. Several considerations taken into account so far are worth noting:

An idea could be very useful and yet be well known by most people, hence fall among those set aside from the first round of screening, where ideas that were deemed "not novel" and/or "not useful" were separated from the rest, leaving the 122 sketches/ideas. Creative ideas often point to new ways of solving engineering/design problems; hence, novel ideas are a better determinant of candidates for creativity compared to usefulness (where incremental improvements often produce very useful ideas). However, usefulness still counts in our assessment. Figure 16 shows the number of ideas that received a 4 (highest score) from a rater, grouped by team for HH and HAI teams.

Figure 16
Highest rating (4, very high) on usefulness and/or novelty for HH and HAI teams



4.8. Statistical tests for significance of differences

To determine the significance of the observed differences above, we applied the independent samples *t*-test on two sets of data. The first one comprised the entire set of 263 ideas while the second used only the highly rated ideas—those which received a "high" or "very high" score from either or both raters. Table 6 shows the results from using the Python function: stats.ttest_ind (HAI, HH, equal_var=False).

Table 6
Statistical tests

Measure	All ideas (263)	Highly rated (122)
t-test statistic	3.814	2.773
p-value	0.00014	0.00729

The means of 3.814 for all ideas and 2.773 for the highly rated ones, as well as the low *p*-values well below the 0.05 threshold, demonstrate confidence in our observations.

This is in addition to evaluating the interrater reliabilities in rating the sketches using two methods—Krippendorff's and ICC—and obtained 0.68 for novelty taken alone, while the combination of usefulness and novelty gave a 0.60 reliability rate. The ICC options were model = "twoway," type = "agreement," unit = "single," and percentage confidence = 95%. Figure 17 shows the output from RStudio.

5. Analysis

We now have sketches ranked on creativity, with each sketch representing a distinct idea. The task (toddler or computer related) is unimportant, so the combined collection is now examined together. 122 sketches from 31 participants remained for evaluation after the first round of grouping by the trained research assistants.

In considering the foregoing points, among others, we determined that focusing on the highest rated ideas worked better than considering all the ideas irrespective of their ratings. Further, by examining these highly rated ideas, we note from Table 7 that:

Figure 17
Statistical tests

```
> kripp.alpha(t(df_nov_use), "ordinal")
 Krippendorff's alpha

 Subjects = 244
   Raters = 2
    alpha = 0.604
> icc(df_nov_use, model = "twoway", type = "agreement", unit = "single")
 Single Score Intraclass Correlation

   Model: twoway
   Type : agreement

   Subjects = 244
     Raters = 2
   ICC(A,1) = 0.603

 F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(243,231) = 4.11 , p = 1.35e-25

 95%-Confidence Interval for ICC Population Values:
  0.516 < ICC < 0.678
```

a) Usefulness and novelty for the 122 sketches

```
> kripp.alpha(t(df_novelty), "ordinal")
 Krippendorff's alpha

 Subjects = 122
   Raters = 2
    alpha = 0.683
> icc(df_novelty, model = "twoway", type = "agreement", unit = "single")
 Single Score Intraclass Correlation

   Model: twoway
   Type : agreement

   Subjects = 122
     Raters = 2
   ICC(A,1) = 0.686

 F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(121,50.9) = 6.02 , p = 4.69e-11

 95%-Confidence Interval for ICC Population Values:
  0.539 < ICC < 0.785
```

b) novelty for the 122 sketches

Table 7
Ideas that received at least one "4" (or highest rating) from the two raters

Sketch#	ASPECT	RATER1	RATER2	TEAM	COMP	TASK
82	Novelty	4	2	3	HH	Toddler
181	Novelty	4	3	8	HH	Toddler
182	Novelty	4	3	8	HH	Toddler
194	Novelty	4	4	9	HH	Toddler
196	Novelty	4	4	9	HH	Toddler
25	Novelty	4	4	1	HAI	Toddler
26	Novelty	4	3	1	HAI	Toddler
126	Novelty	4	4	5	HAI	Toddler
128	Novelty	4	3	5	HAI	Toddler
129	Novelty	4	4	5	HAI	Toddler
146	Novelty	4	3	6	HAI	Toddler
149	Novelty	4	3	6	HAI	Toddler
211	Novelty	4	2	10	HAI	Toddler
212	Novelty	4	4	10	HAI	Toddler
227	Novelty	4	3	11	HAI	Toddler
247	Novelty	4	3	13	HAI	Computer
252	Novelty	4	3	13	HAI	Toddler
405	Novelty	4	3	20	HAI	Toddler
25	Novelty	4	4	1	HAI	Toddler
126	Novelty	4	4	5	HAI	Toddler
129	Novelty	4	4	5	HAI	Toddler
212	Novelty	4	4	10	HAI	Toddler

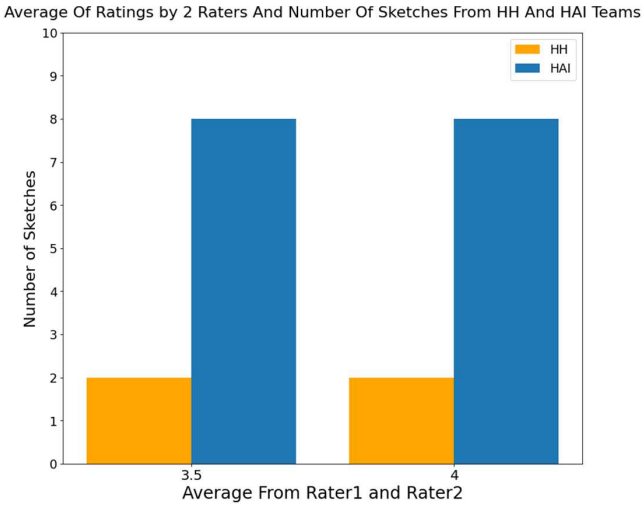
1) 5 ideas by HH teams received a "4" and only 2 got "4" from both raters.
2) 17 ideas by HAI teams received a "4" and 8 got "4" from both raters.

Only 4 sketches received a rating of 4 on "usefulness," and none of them received a 3 or 4 rating on novelty, reinforcing the point made above regarding very useful ideas' tendency to be well known, so likely to be "common." Only 1 "computer" prompt got a rating of 4.

Five ideas by HH teams received a "4," and only 2 got "4" from both raters. Seventeen ideas by HAI teams received a "4," and 8 got "4" from both raters. Table 7 shows all the ideas that received the highest rating of 4 from either or both raters.

Figure 18 represents all the sketches that received the highest rating (4) from both raters, hence an average of 4, or received

Figure 18
HAI teams outperform HH teams



a 3 from one rater and highest (4) from the other rater, hence an average of 3.5. Going by the total number of ideas from the HH and HAI teams, we observe that the HAI teams had 2 ideas for each average (4 total), while the HAI teams had 8 each (16 total).

6. Discussion

Focusing on the statistical significances from the 21 teams highlights the differences in scores from HAI and HH teams, as well as from team to team. This sets up the stage for a discussion of limitations in the study design, as well as alternative study designs that address gaps in the methods that were applied here.

6.1. Quantity of ideas produced by HH and HAI teams

Participants in the HAI teams generated more ideas on average (11.9 sketches) than the HH teams (8.5 sketches) and also had a higher mode of 9, compared with 6 for the HH teams. Given the knowledge that the quantity of ideas matters [54] when it comes to the divergent idea generation stage of the design process, teams would gain an advantage by coming up with as many ideas as possible, thereby expanding the number of candidates for "creative" solutions. It is notable that only 1 out of the 10 HH teams (10%) came up with 15 or more ideas, while 4 of the 11 HH teams (36%) generated over 15 ideas. In addition to this, only 2 HAI teams (18%) had 7 or less ideas compared with 5 (50%) of the HH teams. This difference correlated with the number of high-quality ideas identified across the teams, as will be seen next.

6.2. Quality (usefulness, novelty) of ideas produced by HH and HAI teams

To begin with, 64% of all the 263 sketches received a (lower) score of 1 or 2, while the remaining 36% received a (higher) 3 or 4 rating on usefulness or novelty from either or both of the raters—we count the instances that a 1, 2, 3, or 4 is awarded to a sketch. Any sketch could therefore receive a given score up to 4 times, given that they were rated on usefulness and novelty, and this was done by two raters. The HAI teams obtained higher ratings overall compared with the HH teams. When normalized to account for the difference in the count of ideas from the HH and HAI teams, the HAI teams still had a higher proportion than the HH teams, showing that quantity and quality correlated for Study 2.

6.3. Narrowing down to the highest rated ideas

This step is like separating the signal from the noise of ideas generated by participants. As mentioned earlier, sketches falling into two categories were removed—those from teams that did not follow instructions (by producing complex sketches), as well as one team that was very different from the other participants (both team members knew each other, frequently worked on problem-solving together, and were advanced in their design experience). Of the 216 sketches evaluated after removing these two categories, usefulness received the highest score of 4 from raters 9 times (2%), while novelty received it 29 times (6.7%). Table 8 shows this summary.

Further, we narrowed down to those ideas that received a rating of 4 from both raters or a 4 and a 3 from the raters, given that a 4 and any other rating below 3 would be too varied. This left us with 22 ideas that had an average rating of 4.0 or 3.5 from the 2 raters.

For the 22 ideas rated with a 4.0 or 3.5 average, it is observed that 5 (23%) were by HH teams, while 17 (77%) were from the HAI teams. This represents a large gap in outcomes, as the HAI teams achieved 3 times the number of highly rated ideas compared with the HH teams.

6.4. Collaboration dynamics with the AI

Among the data collected were exit interviews where participants described their experiences and the strategies they had applied in coming up with ideas, as well as in selecting their best idea. These interviews provided some limited insights on how HAI teams interacted with the AI, though the information is insufficient in answering the question as to why these teams performed better than the HH pairs. Below are samples of their responses regarding their experiences, with the first comment alluding to using the AI as inspiration when they ran out of ideas, for example:

"I thought it was pretty interesting. At first, I didn't really think it was going to help me too much besides making my drawings not as terrible, which it did help with because it allowed me to draw a bed somewhat better than how you can usually draw a bed. Where I really found it helpful was in the later stages where

Table 8
Number of times ideas were rated

Aspect	Total evaluated	4 on usefulness	4 on novelty	4 and another 4 or a 3
COUNT/%	216	9 times (2%)	29 times (6.7%)	22 ideas

I was running out of ideas. I would just write a squiggle and I would just scroll through all the suggestions. I'd be like, oh, yeah, we could just do that one."

"And then I carry on with it. Yeah. So without using like pre-made bags or pre-made objects and items. So I would draw something and it would suggest something completely different."

Other ways that HAI participants applied the AI were as a guide in sketching better images, to quickly generate images of interest, and in coming up with completely different suggestions than what they sketched, hence expanding the diversity of their ideas. Nevertheless, a different kind of setup is required to explain the reason for these HAI teams' better performance compared with those who collaborated with another human. One version of this is presented in the next section.

6.5. Relation of results to existing literature

While these results comprise an AI collaborator in early-stage concept development involving novices, there is a rich literature on the role of sketching ideas in supporting creativity. Based on work by Tversky et al. [61] and Simonet et al. [62], some of the reasons HAI collaborators can outperform HH pairs in early-stage sketching tasks lie in the mechanics of perceptual reinterpretation and the mitigation of cognitive fixation. Among these are:

1) Superior management of ambiguity: Tversky et al. [61] argue that the "sketchy" nature of early designs is a deliberate source of creativity. In HH collaboration, two humans often converge too quickly on a single interpretation to maintain social or communicative "consistency."
2) Consensus: an AI is not bound by the human need for social consensus. It can generate multiple, divergent "functional inferences" from a single vague sketch. By presenting "reinterpreted" versions of a human's sketch that the human did not initially intend, the AI acts as a catalyst for what Tversky et al. [61] call "new perceptual groupings," pushing the human designer beyond their initial mental model.
3) Bridging the functional-perceptual gap: an AI can be trained specifically to bridge this gap. While a human designer (especially a novice) might be stuck looking at the "lines" of a sketch, an AI can instantly overlay functional possibilities onto the sketchy configuration. This provides a "novice" human with "expert-level" functional insights in real time.
4) Relief of working memory and "the seeing-moving-seeing" cycle: sketching serves to relieve short-term memory [61]. In HH collaboration, the cognitive load of communicating ideas to another person can actually compete with the cognitive load of the design task itself. HAI interaction is often more "fluid" and less socially demanding than HH interaction. The AI can rapidly iterate on the "ambiguous groupings" of a sketch, allowing the human to stay in a continuous loop of "reinterpretation." Because the AI can process and "see" new relations in the sketch faster than a human partner, it accelerates the cycle of making new inferences, leading to a higher volume of creative breakthroughs.

Expanding on this, recent studies by Figoli et al. [63] suggest that AI serves as a superior "empathy trigger" and visual stimulus. Given that the AI doesn't "know" the designer's intent, its "misinterpretations" of a human sketch act as powerful external stimuli that force the human to see new functional groupings they would otherwise ignore.

6.6. Implications for scalable evaluation

The use of an AI-based image classifier to evaluate sketches marks a significant shift toward automated assessment in educational settings. An example of this is in early concept development in product design courses, where a few dozen students may sketch multiple versions of their ideas, leading to hundreds or even thousands of sketches produced in a 3-h session. With a system that enables students to submit their sketches in real time, this classifier can select "high quality" or "novel" ideas—based on some set criteria—and present them to the instructor to gauge how well students are learning. In addition to this, using no-code tools like Teachable Machine makes this high-level technology accessible to educators who may not have a background in computer science. By integrating human expertise with ML efficiency, this experimental tool demonstrates a methodology capable of managing the high-volume output of modern engineering classrooms. Table 9 presents the comparison between human and AI-assisted evaluation.

6.7. Approach Limitations

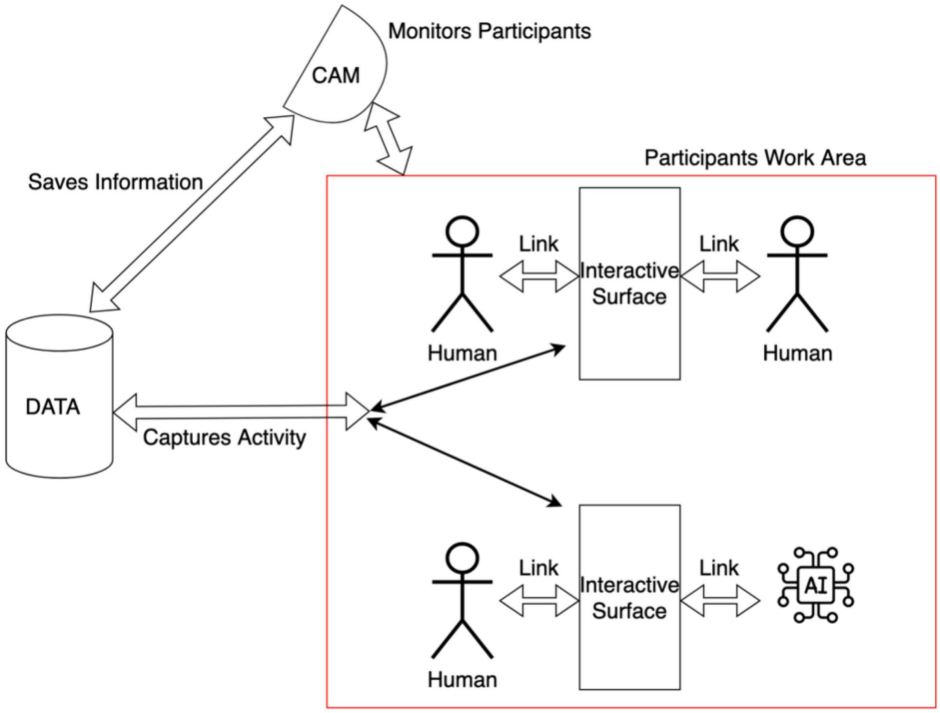
The first limitation arises from the manner in which HH collaborated relative to the HAI teams—the humans sat next to each other and occupied the same space during the activity while speaking, gesturing, and sketching in turns, while the HAI teams had a human work with an AI that interacted through a screen by displaying images. This raises the possibility of confounding factors such as:

1) Humans in HH teams had to wait while listening to each other, while the HAI teams could simultaneously proceed without turn-taking, introducing a delay in moving to the next idea. On the flip side, two humans might have an advantage of combined experience relative to a human collaborating with an AI that has little experience of the world.
2) HH teams had a wider range of means to express themselves, both verbal and nonverbal, while the HAI teams relied on the AI displaying images alone, limiting how they perceived and responded to information. Traberg et al. [64] have suggested that such perception among teammates impacts creativity. However, limiting the communication mode may have helped

Table 9
Human vs AI-assisted evaluation

Feature	Traditional human evaluation	AI-assisted evaluation (CNN/TM)
Speed	Slow (limited by human hours)	Instant (real-time classification)
Scalability	Low (difficult with 100+ students)	High (handles thousands of sketches)
Consistency	Subject to fatigue and bias	100% consistent based on training
Flexibility	High (can adapt to nuances)	Moderate (requires retraining for new rubrics)

Figure 19
Alternative configuration for capturing moment-to-moment interactions



focus the humans in the HAI teams, giving them an advantage over the HH teams.

3) HH teams both knew the context of the task, something that an AI lacked—which could be an advantage or disadvantage on divergent ideation for the HH teams. A possibility with the HAI teams and their higher number of ideas relative to HH teams could be that the AI, lacking context, provided better and new suggestions for leading the human to novel solutions.

Another identified limitation is that we lack an understanding of how the AI-generated suggestions impacted human collaborators as they conceived ideas to solve the task. The humans had the option to use the suggested ideas (sketches) in the following ways:

1) Take it as is, thereby suggesting that the idea was deemed appropriate and sufficient.
2) Take and modify the idea, which would indicate that the human found the idea to be relevant but also incomplete or insufficient without the additional details they provide.
3) Ignore the suggestion, suggesting that it was deemed inappropriate, irrelevant, or for another reason.
4) Ignore the suggestion, but use it as an inspiration for a new idea, as would be the case of a shift in paradigm that some participants expressed in the exit interviews.

Given the nature of our experimental setup, this study focuses on the outcomes of HAI collaboration rather than capturing the real-time dynamics of interaction, which might have yielded richer behavioral data. While we acknowledge this as a limitation, prior research such as the work of Shah et al. [39] and others has effectively assessed idea quality based solely on sketched outputs, supporting the validity of our approach. Additionally, participant debriefing occurred afterward, requiring retrospective recall rather than reflecting on their collaboration as it unfolded. This reliance on memory may limit the granularity of insight into the interaction process. Future studies could address this limitation by incorporating video observation and think-aloud protocols, enabling a more nuanced understanding of in-the-moment cognitive and collaborative processes.

One such design that captures the moment-to-moment interactions—as well as if, when, and how the human uses the AI suggestions—is shown in Figure 19, where a camera (CAM) captures live video of the participants interacting with the device on which the AI is running, while a separate system captures the device activity.

7. Summary

This paper explored how AI can enhance idea generation during the concept development phase of engineering design. We introduced the concept of a "Disruptive Interjector (DI)"—an AI system that continuously observes a user's actions and interjects with contextually relevant suggestions to stimulate new directions in creative thinking. Both expert human raters and a CNN-based classifier evaluated ideas for novelty and usefulness, which demonstrated that HAI teams consistently outperformed or matched human-only teams by yielding a higher proportion of creative outcomes and a greater volume of ideas. This suggests that AI, when framed as a disruptive partner, can effectively support divergent thinking by introducing novel stimuli that push designers toward new directions. These results provide empirical evidence that AI can transcend the role of a passive tool to become an active, superior collaborator in creative tasks. Furthermore, the study introduces a scalable, replicable method for high-volume evaluation in educational settings, proving that AI-assisted assessment can align with expert human judgment to manage the vast datasets typical of design studios and classrooms.

As AI technologies increasingly become integrated into creative and technical domains, this study demonstrates the efficacy of a class of AI systems that, rather than providing solutions to prompts, offers suggestions that serve as inspiration for new ideas. This is a potential starting point for building effective AI-based systems that support human creativity.

Ethical Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of STANFORD UNIVERSITY (eProtocol # 50409 approved on 24/05/2019) for studies involving humans.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://github.com/makokhamj/H-AI-Study.

Author Contribution Statement

Joseph M. Makokha: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

References

[1] Kuo, R. Y. L., Harrison, C., Curran, T.-A., Jones, B., Freethy, A., Cussons, D., . . . , & Furniss, D. (2022). Artificial intelligence in fracture detection: A systematic review and meta-analysis. *Radiology*, *304*(1), 50–62. https://doi.org/10.1148/radiol.211785

[2] Cortellini, A., Santo, V., Brunetti, L., Garbo, E., Pinato, D. J., la, Cava., G, Guarrasi, V., . . . , Guarrasi, V. (2025). Transformer-based AI approach to unravel long-term, time-dependent prognostic complexity in patients with advanced NSCLC and PD-L1 ≥50%: Insights from the pembrolizumab 5-year global registry. *Journal for Immunotherapy of Cancer*, *13*(9), e012423. https://doi.org/10.1136/jitc-2025-012423

[3] Varghese, C., Harrison, E. M., O'Grady, G., & Topol, E. J. (2024). Artificial intelligence in surgery. *Nature Medicine*, *30*(5), 1257–1268. https://doi.org/10.1038/s41591-024-02970-3[3]

[4] Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *18*(2), 66. https://doi.org/10.1145/3475799

[5] Goumiri, S., Yahiaoui, S., & Djahel, S. (2025). Smart mobility in smart cities: Emerging challenges, recent advances and future directions. *Journal of Intelligent Transportation Systems*, *29*(1), 81–117. https://doi.org/10.1080/15472450.2023.2245750

[6] Zhang, S., Wang, H., & Yi, X. (2025). Exploring collaboration patterns and strategies in human-ai co-creation through the lens of agency: A scoping review of the top-tier HCI literature. *arXiv*. *Preprint:2507.06000*

[7] Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., . . . , & Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074. https://doi.org/10.1126/science.ade9097

[8] Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., . . . , & Gaggioli, A. (2023). Artificial intelligence & creativity: A manifesto for collaboration. *The Journal of Creative Behavior*, *57*(4), 472–484. https://doi.org/10.1002/jocb.597

[9] Makokha, J. M. (2022). *Collaborative Artificial Intelligence (AI) for idea generation in design teams*. Stanford University.

[10] Ceh, S. M., Rafner, J., & Benedek, M. (2024). Creativity in digitally mediated times: How digital tools support creativity across domains. *Psychology of Aesthetics, Creativity, and the Arts*, *20*(3), 794–808. https://doi.org/10.1037/aca0000729

[11] Rezwana, J., & Maher, M. L. (2023). Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction*, *30*(5), 67. https://doi.org/10.1145/3519026

[12] Kempkes, J. A., Suprano, F., & Wömpener, A. (2023). How management support systems affect job performance: A systematic literature review and research agenda. *Management Review Quarterly*, *74*(4-5), 1–74. https://doi.org/10.1007/s11301-023-00353-5

[13] Tang, C., Mao, S., Naumann, S. E., & Xing, Z. (2022). Improving student creativity through digital technology products: A literature review. *Thinking Skills and Creativity*, *44*, 101032. https://doi.org/10.1016/j.tsc.2022.101032

[14] Sinnemann, M. F., & Weiss, M. M. (2025). Team virtuality and innovation: A meta-analysis of the moderating role of team design. *Journal of Organizational Behavior*, *46*(6), 867–888. https://doi.org/10.1002/job.2889

[15] Abi Saad, E., & Agogué, M. (2023). Creativity in virtual teams: Systematic review, synthesis and research agenda. *Creativity and Innovation Management*, *32*(1), 117–140. https://doi.org/10.1111/caim.12540

[16] Langley, P. (2025). Learning hierarchical task knowledge for planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(27), 28652–28656. https://doi.org/10.1609/aaai.v39i27.35091

[17] Baya, V., Cannon, D. M., Lakin, F., Baudin, C., & Leifer, L. (1990). *The electronic design notebook*. Stanford University, Center for Design Research. http://www-cdr.stanford.edu/html/GCDK/vmacs/edn.html

[18] Yun, G., Cho, K., Jeong, Y., & Nam, T.-J. (2022). Ideasquares: Utilizing generative text as a source of design inspiration. In *Proceedings of the Design Research Society International Conference 2022*, 1–20. https://doi.org/10.21606/drs.2022.484

[19] Auernhammer, J., Leifer, L., Meinel, C., & Roth, B. (2022). A humanistic and creative philosophy of design. In *Design thinking research: Achieving real innovation* (pp. 1–15). https://doi.org/10.1007/978-3-031-09297-8_1

[20] Dow, G. T. (2022). Defining creativity. In J. A. Plucker (Ed.), *Creativity and innovation* (pp. 5–21). Routledge. https://doi.org/10.4324/9781003233923

[21] Runco, M. A. (2023). AI can only produce artificial creativity. *Journal of Creativity*, *33*(3), 100063. https://doi.org/10.1016/j.yjoc.2023.100063

[22] Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1). 4–11. https://doi.org/10.1109/THFE2.1960.4503259

[23] Simon, H. A., & Newell, A. (1958). Heuristic problem solving: The next advance in operations research. *Operations Research*, *6*(1), 1–10. https://doi.org/10.1287/opre.6.1.1

[24] Martí, R., Sevaux, M., & Sörensen, K. (2025). Fifty years of metaheuristics. *European Journal of Operational Research*, *321*(2), 345–362. https://doi.org/10.1016/j.ejor.2024.04.004

[25] Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators. (Report No. ADA057655)*. Massachusetts Institute of Technology, Man-Machine Systems Laboratory. https://doi.org/10.21236/ADA057655

[26] Richardson, L. S., Fidock, J., & Gunawan, I. (2025). Systematic literature review of levels of automation (autonomy) taxonomy: Critiques and recommendations. *International Journal of Human–Computer Interaction*, *41*(24), 15824–15843. https://doi.org/10.1080/10447318.2025.2502978

[27] Moruzzi, C., & Margarido, S. (2024). A user-centered framework for human-AI co-creativity. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3613905.3650929

[28] Fisher, C. M., Pillemer, J., & Amabile, T. M. (2025). When the thought doesn't count: The dynamics of unhelpful help in creative organizations. *Academy of Management Discoveries*, *11*(3), 342–357. https://doi.org/10.5465/amd.2023.0160

[29] Fletcher, A., & Benveniste, M. (2022). A new method for training creativity: Narrative as an alternative to divergent thinking. *Annals of the New York Academy of Sciences*, *1512*(1), 29–45. https://doi.org/10.1111/nyas.14763

[30] Lee, J. H., & Ostwald, M. J. (2022). The relationship between divergent thinking and ideation in the conceptual design process. *Design Studies*, *79*, 101089. https://doi.org/10.1016/j.destud.2022.101089

[31] Herault, C., Ovando-Tellez, M., Lebuda, I., Kenett, Y. N., Beranger, B., Benedek, M., & Volle, E. (2024). Creative connections: The neural correlates of semantic relatedness are associated with creativity. *Communications Biology*, *7*(1), 810. https://doi.org/10.1038/s42003-024-06493-y

[32] Childs, P., Han, J., Chen, L., Jiang, P., Wang, P., Park, D., . . . , & Vilanova, I. (2022). The creativity diamond—A framework to aid creativity. *Journal of Intelligence*, *10*(4), 73. https://doi.org/10.3390/jintelligence10040073

[33] Benedek, M. (2024). On the relationship between creative potential and creative achievement: Challenges and future directions. *Learning and Individual Differences*, *110*, 102424. https://doi.org/10.1016/j.lindif.2024.102424

[34] Eluchans, M., Lancia, G. L., Maselli, A., D'Alessandro, M., Gordon, J. R., & Pezzulo, G. (2025). Adaptive planning depth in human problem-solving. *Royal Society Open Science*, *12*(4), 241161. https://doi.org/10.1098/rsos.241161

[35] Abraham, A. (2025). Why the standard definition of creativity fails to capture the creative act. *Theory & Psychology*, *35*(1), 40–60. https://doi.org/10.1177/09593543241290232

[36] Sternberg, R. J. (2022). The field of psychology never maxed out on the ideas of Max Wertheimer: A new look at productive thinking. *The American Journal of Psychology*, *135*(2), 248–250. https://doi.org/10.5406/19398298.135.2.12

[37] Guilford, J. P. (1950). Creativity. *American Psychologist*, *5*(9), 444–454. https://doi.org/10.1037/h0063487

[38] Bruner, J. S. (1962). The conditions of creativity. In H. E. Gruber, G. Terrell, & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking: A symposium held at the University of Colorado* (pp. 1–30). Atherton Press. https://doi.org/10.1037/13117-001

[39] Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design Studies*, *24*(2), 111–134. https://doi.org/10.1016/S0142-694X(02)00034-0

[40] Fiorineschi, L., & Rotini, F. (2023). Uses of the novelty metrics proposed by Shah et al.: What emerges from the literature? *Design Science*, *9*, e11. https://doi.org/10.1017/dsj.2023.9

[41] Pichot, N., Bonetto, E., Pavani, J.-B., Arciszewski, T., Bonnardel, N., & Weisberg, R. W. (2022). The construct validity of creativity: Empirical arguments in favor of novelty as the basis for creativity. *Creativity Research Journal*, *34*(1), 2–13. https://doi.org/10.1080/10400419.2021.1997176

[42] Peeperkorn, M. (2022). Artificial creative societies: Adaption, intention, and evaluation. In *Proceedings of the 14th Conference on Creativity and Cognition*, 704–707. https://doi.org/10.1145/3527927.3533728

[43] Simonton, D. K. (2012). Taking the U.S. Patent Office creativity criteria seriously: A quantitative three-criterion definition and its implications. *Creativity Research Journal*, *24*(2-3), 97–106. https://doi.org/10.1080/10400419.2012.676974

[44] Cascini, G., Nagai, Y., Georgiev, G. V., Zelaya, J., Becattini, N., Boujut, J. F., . . . , & Wodehouse, A. (2022). Perspectives on design creativity and innovation research: 10 years later. *International Journal of Design Creativity and Innovation*, *10*(1), 1–30. https://doi.org/10.1080/21650349.2022.2021480

[45] Gero, J., & Milovanovic, J. (2024). Do creativity metrics from design research correlate with those from psychology? *Creativity Research Journal*, *36*(3), 508–520. https://doi.org/10.1080/10400419.2024.2320513

[46] Weisberg, R. W. (2015). On the usefulness of "value" in the definition of creativity. *Creativity Research Journal*, *27*(2), 111–124. https://doi.org/10.1080/10400419.2015.1030320

[47] Oppert, M. L., O'Keeffe, V., Bensnes, M. S., Grecu, A. L., & Cropley, D. H. (2023). The value of creativity: A scoping review. *Journal of Creativity*, *33*(2), 100059. https://doi.org/10.1016/j.yjoc.2023.100059

[48] Barron, F. (1955). The disposition toward originality. *The Journal of Abnormal and Social Psychology*, *51*(3), 478–485. https://doi.org/10.1037/h0048073

[49] Newell, A., Shaw, J. C., & Simon, H. A. (1962). The processes of creative thinking. In H. E. Gruber, G. Terrell, & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking: A symposium held at the University of Colorado* (pp. 63–119). Atherton Press. https://psycnet.apa.org/doi/10.1037/13117-003

[50] Murphy, L. R., Daly, S. R., & Seifert, C. M. (2023). Idea characteristics arising from individual brainstorming and design heuristics ideation methods. *International Journal of Technology and Design Education*, *33*(2), 337–378. https://doi.org/10.1007/s10798-021-09723-0

[51] Makokha, J. M. (2023). Augmenting people's creative idea generation using an artificial intelligent sketching collaborator. *International Journal of Computer and Information Engineering*, *17*(1), 84–92.

[52] Wang, S., Okada, T., & Takagi, K. (2023). How to effectively overcome fixation: A systematic review of fixation and defixation studies on the basis of fixation source and problem type. In *Frontiers in Education*, *8*, 1183025. https://doi.org/10.3389/feduc.2023.1183025

[53] Sreenivasan, A., & Suresh, M. (2024). Design thinking and artificial intelligence: A systematic literature review exploring synergies. *International Journal of Innovation Studies*, *8*(3), 297–312. https://doi.org/10.1016/j.ijis.2024.05.001

[54] Nguyen, M., & Mougenot, C. (2022). A systematic review of empirical studies on multidisciplinary design collaboration: Findings, methods, and challenges. *Design Studies*, *81*, 101120. https://doi.org/10.1016/j.destud.2022.101120

[55] Zhang, W. (2024). Ebb and flow: Design fixation and creativity in professional groups. *Journal of Engineering Design*, *35*(3), 263–289. https://doi.org/10.1080/09544828.2024.2306783

[56] Sahaai, M. B., Jothilakshmi, G. R., Ravikumar, D., Prasath, R., & Singh, S. (2022). ResNet-50 based deep neural network using transfer learning for brain tumor classification. In *AIP Conference Proceedings*, *2463*(1), 020014. https://doi.org/10.1063/5.0082328

[57] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, *57*(4), 99. https://doi.org/10.1007/s10462-024-10721-6

[58] Yu, X., Wang, J., Hong, Q.-Q., Teku, R., Wang, S.-H., & Zhang, Y.-D. (2022). Transfer learning for medical images analyses: A survey. *Neurocomputing*, *489*, 230–254. https://doi.org/10.1016/j.neucom.2021.08.159

[59] Lloyd-Cox, J., Pickering, A., & Bhattacharya, J. (2022). Evaluating creativity: How idea context and rater personality affect considerations of novelty and usefulness. *Creativity Research Journal*, *34*(4), 373–390. https://doi.org/10.1080/10400419.2022.2125721

[60] Cross, N. (2024). Creative cognition in design II: Co-evolution of problem and solution. In *Designerly Ways of Knowing and Thinking* (pp. 63–73). https://doi.org/10.1007/978-1-4471-7541-4_5

[61] Tversky, B., Suwa, M., Agrawala, M., Heiser, J., Stolte, C., Hanrahan, P., . . . , & Haymaker, J. (2003). Sketches for design and design of sketches. In U. Lindemann (Ed.), *Human behaviour in design* (pp. 79–86). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-07811-2_9

[62] Simonet, M., Vater, C., Abati, C., Zhong, S., Mavros, P., Schwering, A., . . . , & Krukar, J. (2025). Probing mental representations of space through sketch mapping: A scoping review. *Cognitive Research: Principles and Implications*, *10*(1), 59. https://doi.org/10.1186/s41235-025-00667-w

[63] Figoli, F. A., Rampino, L., & Mattioli, F. (2022). AI in design idea development: A workshop on creativity and human-AI collaboration. In *Proceedings of the Design Research Society International Conference 2022*, 1–17. https://doi.org/10.21606/drs.2022.414

[64] Traberg, C. S., Harjani, T., Roozenbeek, J., & Van der Linden, S. (2024). The persuasive effects of social cues and source effects on misinformation susceptibility. *Scientific Reports*, *14*(1), 4205. https://doi.org/10.1038/s41598-024-54030-y

How to Cite: Makokha, J. M. (2026). Future AI Systems: Human–AI Collaborative Teams Outperform Human-Only Teams in Design Ideation. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027260