Artificial Intelligence for Engineering Design, Analysis and Manufacturing

www.cambridge.org/aie

Creative potential of image-generative AI models for conceptual engineering design tasks

Vicente Chulvi[1], Laura Ruiz-Pastor[2], Aurora Berni[3], Marta Royo[1] and Ana Castillo-López[1]

[1]Enginyeria Mecànica i Construcció, University Jaume I, Spain; [2]Industrial Systems Engineering and Design, University Jaume I, Spain and [3]Free University of Bozen-Bolzano, Italy

Research Article

Cite this article: Chulvi V, Ruiz-Pastor L, Berni A, Royo M and Castillo-López A (2026). Creative potential of image-generative AI models for conceptual engineering design tasks. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **40**, e8, 1–15
https://doi.org/10.1017/S0890060426100286

Received: 06 March 2026 UTC
Revised: 03 July 2026 UTC
Accepted: 15 August 2026 UTC

Keywords:
creativity; creative design; idea generation; artificial intelligence; image-generative AI

Corresponding author:
Vicente Chulvi;
Email: chulvi@uji.es

Abstract

Creativity is paramount in design engineering, driving design innovation with new ideas. Visual representations are crucial in communicating and refining design ideas. The emergence of image-generative AI presents new opportunities, but a comprehensive understanding of AI creative capabilities, particularly in comparison to human designers, remains unexplored. This explorative study investigates the creative potential of image-generative AI systems by comparing their outputs to those of human designers. Utilizing the metrics of Quantity, Variety, Novelty, and Quality, we evaluated the performance of 6 AI-based image-generation tools representing several diffusion-model families and 20 design engineering students on an assigned design task. Human participants were subject to time constraints, while AI tools were constrained by predefined output parameters. Our findings highlight the complementary nature of human and AI creativity. AI tools excel at generating a large volume of solutions rapidly, but their outputs lack the diversity, novelty, and nuanced quality characteristic of human designs. While AI tools demonstrated promising results in Novelty, achieving high Quality consistently proved challenging across all tools. These results underscore the importance of integrating human oversight and interpretation to guide AI systems towards more meaningful and innovative outcomes, with the aim of optimizing the process of generating creative ideas of designers, using AI systems for image generation.

Introduction

Creativity is a fundamental aspect in design engineering, serving as a critical driver for generating innovative concepts and solving problems. It involves a mental process of producing new and useful ideas, often linked to the ability to think divergently – the capability to develop multiple solutions to a given problem (Guilford, 1950). According to Amabile (1983), creativity is a cognitive process encompassing the generation of original and practical ideas, combining existing concepts, and transforming known ideas into novel forms. Traditionally, creativity has been seen as a uniquely human capability (Amabile, 2019). However, the rise of artificial intelligence (AI) has challenged this view, spurring interest among companies and researchers in understanding the potential of AI to contribute to innovation processes (Haefner et al., 2021).

The rapid development of AI models has expanded their use as a method for supporting creative tasks in design engineering (Oktradiksa et al., 2021). AI models are increasingly seen not just as functional aids but as tools capable of influencing creativity itself (Chulvi, 2025). While text-based AI models, such as ChatGPT, have been studied extensively for their impact on creative thinking in product design (e.g., Cropley, 2023; Filippi, 2023), less attention has been given to the creative potential of image-generative AIs. These systems interpret specific prompts and transform them into visual representations, offering new possibilities for visually-driven design tasks.

Theoretical and practical opportunities for AI in design engineering are immense. Theoretically, researchers are investigating how AI models can complement the design process (Bouschery et al., 2023) while practically, new AI models are being developed to generate novel and diverse design concepts (Dai et al., 2023). However, a significant gap exists in the literature: most studies explore AI as a collaborative partner for human designers without fully comparing the creative outputs of AI and humans first. This raises a critical issue of understanding the advancements of image-generative AI models in terms of creativity while performing design tasks.

Moreover, while some studies involve the use of text-generative AI models such as ChatGPT, the potential of image-generative AI systems to produce creative and impactful visual outputs in design engineering remains underexplored, despite visual representations playing a vital role in creative product design, often extending beyond the scope of textual descriptions (Townsend and Kahn, 2014).

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

The objective of this article is to evaluate the potential of image-generative AI systems to support designers during the creative idea-generation phase. Specifically, the study compares the creativity of visual outputs produced by humans and image-generative AI tools.

To achieve this, an experiment was conducted where both human participants and image-generative AI tools were tasked with addressing the same design problem. Participants were instructed to produce as many ideas as possible within a 20-minute timeframe. The resulting outputs were evaluated using the four parameters of creativity proposed by Shah et al. (2003): quantity, variety, novelty, and quality. This evaluation framework enables a rigorous comparison between human and AI-generated outputs, as well as an analysis of differences across multiple AI tools. The experimental details are provided in section "Preparation of the experiment."

The paper is structured as follows: Section "Background" defines the background providing a literature review on the use of AI for creativity, stressing the importance of visual stimuli in design and creativity (2.1) and the potential of image-generative AIs in generating creative design concepts. To answer the RQs that emerged from the literature (see section "Background") the experimental procedure described in section "Methodology" was designed. Results of the analysis can be found in section "Results" followed by their discussion in section "Discussion." Conclusions are presented in section "Conclusion."

Background

Visual representation in creative design

Design – whether in engineering, industrial and product design, graphic design, or architecture – relies heavily on visual outputs. These outputs, often referred to as visual representations, may be textual, static, or dynamic (Pei et al., 2011; Berni et al., 2020). Designers frequently use such representations to communicate ideas and present concepts effectively. Among these, visual static representations play a critical role, serving as stimuli that evoke emotions, influence decision-making, and shape thought processes (Borgianni et al., 2020). Using visual stimuli is particularly significant during the early ideation stages of the design process, where creativity and exploration are paramount.

In design creativity, visual stimuli play a critical role, serving as both a source of inspiration and a potential limitation.

Borgianni et al. (2020) underscored the importance of external sources of inspiration, particularly incorporating both text-based and graphic forms of stimuli, noting that intentionally selected external inputs can significantly enrich designers' creative processes.

Agogué et al. (2014) confirm that the type of stimuli provided plays a critical role in shaping creative idea generation during problem-solving tasks. The scholars investigated the impact of text-based stimuli on creative flow, specifically how such examples can lead to design fixation. They identify two types of examples: "restrictive examples," which limit possibilities by activating common knowledge and thereby reducing originality, and "expansive examples," which introduce unexpected ideas, fostering novel solutions and enhancing originality.

Similarly, research by Cardoso and Badke-Schaub (2011) highlights the dual nature of visual inputs: while they can spark innovative ideas, they also risk inducing design fixation, where designers become overly focused on specific attributes of a presented visual, reducing the diversity of their outputs. This phenomenon was particularly evident in their study involving industrial design students, where participants exposed to certain pictorial representations showed higher levels of repetitive patterns compared to those in a control group. Such findings emphasize the need for varied and carefully curated visual stimuli to mitigate fixation and expand creative exploration. Based on the contributions above, visual stimuli, such as images, emerge to be particularly effective due to their ability to evoke emotions and facilitate connections that might not occur otherwise. The scholars suggest that diverse visual prompts can encourage lateral thinking, enabling designers to explore unconventional ideas. In recent years, while traditional visual stimuli have long played a pivotal role in design, the emergence of AI – particularly generative AI – has revolutionized their creation and application. This transformation presents new opportunities and challenges that remain to be fully explored, including whether AI can mitigate or exacerbate fixation compared to traditional stimuli, and whether AI models themselves may be susceptible to fostering creative fixation.

Generating AI for creative design: opportunities and challenges

Artificial Intelligence (AI) has experienced rapid growth, revolutionizing traditional workflows across many fields and industries (Bordas et al., 2024). Among the different AI branches, generative AI has gained prominence due to its ability to produce diverse and high-quality outputs such as text, images, videos, and even 3D designs (Kılınç and Keçecioğlu, 2024). Unlike AIs designed for prediction or classification tasks, generative AI excels in creating novel content by learning from patterns in large datasets (Bordas et al., 2024; Kılınç and Keçecioğlu, 2024). Its versatility is reflected in the variety of input–output that ranges from text-to-text, text-to-code to text-to-images etcetera, expanding AI application from coding, writing to generating images and multimedia outputs (Kılınç and Keçecioğlu, 2024). Tools like DALL-E, MidJourney, and Stable Diffusion leverage these models to produce visually realistic works, potentially fostering creativity and iterative design processes in fields such as art, engineering, and architecture Hutson & Cotroneo (2023). Furthermore, generative AI's capability to rapidly generate content, whether in written or visual forms, has made it an invaluable asset for both creators and industries. However, the limits and potential of generative AI are still debated; thus, further investigation is needed (Bordas et al., 2024). The adaptability of generative AI is particularly evident in its application to text-based and image-based creative tasks, each with distinct capabilities and challenges.

Text-to-text AI (Text-generative AI)

Text-generative AI models, such as OpenAI's ChatGPT, have demonstrated significant potential in enhancing idea generation for creative design (Wang et al.,2025a). These systems are capable of producing written content, brainstorming concepts, and simulating human-like conversations based on user prompts Uludag (2023). In product design, for instance, Filippi (2023) identifies certain challenges, including the risk of reduced creativity and the difficulty of translating generated ideas into tangible designs. In a study assessing ChatGPT's impact on innovative concept generation in product design, the scholar validated hypotheses related to four specific metrics: quantity, usefulness, novelty, and variety. The study involved 18 university students and revealed that ChatGPT significantly increased the quantity and novelty of generated concepts, although its contribution in terms of usefulness and variety was less pronounced. Surprisingly, despite ChatGPT's reliance on existing content, the greater quantity of concepts generated was associated with increased novelty. In the same line, Chulvi et al.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

(2025) lead a GPT-4-based generative AI to use a design methodology to generate creative concepts, were results suggest that the solutions provided by the AI when applying the SCAMPER design methodology are as creative as the average of the students using the same methodology. Additionally, Stock-Homburg (2023) conducted a study involving 689 participants, asking them to assess the perceived novelty of numerous innovations and classify them based on whether they believed they were generated by a human or an AI (specifically, ChatGPT). Surprisingly, her findings revealed that participants were unable to distinguish the origin of the innovations presented, regardless of their experience with ChatGPT or familiarity with innovations. Nevertheless, based on results from a study investigating ChatGPT's verbal divergent thinking ability, Cropley (2023) suggests that the most promising application of AI in creative tasks lies in supporting human divergent thinking. In a similar vein, Doshi and Hauser (2023) demonstrated that the use of generative AI enhances the creativity of writers by up to 11% compared to human writers without AI assistance. Additionally, Noy and Zhang (2023) found that generative AI reduces the time required for writing by up to 20% and significantly improves the quality of work produced by low-ability workers, achieving an enhancement of up to 40%.

Text-to-image AI (image-generative AI)

Image-generative AI models, such as DALL-E, Stable Diffusion, and MidJourney, have emerged as potential tools for creative design, capable of producing realistic and visually compelling images from text-based inputs (Chen and Chen, 2023). These systems leverage neural networks trained on extensive datasets to generate diverse visual output, fundamentally reshaping the traditional idea generation process. By offering new creative possibilities, these tools influence aesthetics and alter the ways creators conceptualize and produce art and design concepts (Epstein et al., 2023).

By allowing designers to visualize abstract ideas and refine concepts dynamically, these tools hold the potential to foster creativity, improve workflow efficiency, and encourage novel concept generation in engineering design (Brisco et al., 2023).

However, recent studies highlight both the promise and limitations of image-generative AI models and their outputs. Mhlanga (2024) and Fui-Hoon Nah et al. (2023) highlight critical risks associated with image-generative AI misuse and over-reliance:

- Ethically, AI models risk perpetuating biases embedded in their training data, potentially reinforcing discrimination or generating inappropriate content.
- Technologically, the quality and representativeness of AI-generated outputs rely heavily on the underlying training data, raising concerns about reliability. Consequently, the result is also strongly dependent on the prompt formulated, since slightly different prompts could lead to substantially different results.
- As AI-generated content becomes increasingly sophisticated, the boundary between human and AI-created work is blurring. While AI models still have limitations, such as producing unrealistic or nonspecific content (Brisco et al., 2023), their ability to mimic human creativity raises concerns about the role of human designers in the creative process. The rapid generation of diverse, high-quality outputs challenges the notion of AI as a mere tool and raises questions about its potential to replace human designers or foster genuine creativity.

Research into the interaction between human designers and generative AI models underscores the complexity of these issues. Turchi et al. (2023) highlight the growing adoption of image-generative AI models but emphasize the need for further research on their impact on creativity and designer control. Wang et al. (2023) demonstrate DALL-E's capacity for combinational creativity, approaching human levels in visualizing hybrid ideas. Wang and Han (2023) further reveal the positive influence of generated visual stimuli on design ideation, particularly among experienced designers.

While these studies showcase the potential of image-generative AI in specific design contexts (Ishihara et al., 2023; Paananen et al., 2024; Ren and Du, 2024; Zhang et al., 2024), they often focus on domain-specific applications or the appropriateness of outputs. A more comprehensive understanding of AI's creative potential requires systematic comparisons with human creativity. By delving deeper into the nuances of human-AI collaboration and the unique contributions of each, we can better harness the power of AI to augment human creativity rather than replace it. Recent research has highlighted the growing role of generative AI as a creativity support tool in conceptual design processes (Chen et al., 2025). However, most existing studies focus on human–AI collaboration dynamics rather than on direct comparisons between independently generated AI concepts and human-generated design outcomes. Although comparative studies have recently emerged in the context of human–AI co-creation (Wang et al., 2025b), empirical evidence remains limited regarding the standalone creative potential of image-generative AI systems in conceptual engineering design tasks.

Gaps and rationale of the study

Although prior studies have explored AI-assisted creativity, creative ideation, and human–AI co-creation workflows, comparatively fewer studies have conducted controlled empirical comparisons between image-generative AI systems and human designers specifically within conceptual engineering design tasks, and particularly during early-stage ideation. Moreover, existing comparative studies are frequently situated in graphic design, artistic production, or co-creative workflows, leaving conceptual engineering design comparatively underexplored. On the field of engineering design, this gap raises critical questions about whether AI can rival or surpass human creativity or whether it primarily serves as a complementary tool. Therefore, the gap is not the absence of research on AI and creativity per se, but rather the limited number of studies examining the independent creative potential of image-generative AI systems in engineering-oriented conceptual design contexts.

To address this gap, the study formulates two research questions:

- RQ1: How do image-generative AI outputs differ from human design outputs in terms of creativity?
- RQ2: How do different image-generative AI systems vary in their ability to improve creativity, and which creativity dimensions differ the most across AI systems?

Understanding the degree of creativity demonstrated by image-generative AI in producing design concepts is essential for evaluating its role in creative workflows. RQ1 focuses on comparing AI-generated and human-generated outputs in identifying strengths and weaknesses of AI tools in specific creativity metrics, such as novelty, quantity, and variety (Shah et al., 2003).

Additionally, while tools like DALL-E and Stable Diffusion share similar foundational models, differences in their training processes result in varying output styles, specifications, and levels of realism. These differences may impact their creative capabilities, leading to the formulation of RQ2. These open questions underscore the need

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

for an in-depth investigation into the creative potential of image-generative AI systems, particularly in direct comparison with human designers. This study aims to address these gaps by examining the nature of AI creativity and proposing a framework for evaluating the quality and originality of AI-generated concepts.

To address these research questions, the authors conducted an experiment in which both humans and image-generative AI tools were given a design task where they had to generate ideas for a specific design problem. Human participants were instructed to produce as many ideas as possible, with a time limit of 20 minutes. The creativity of the resulting visual outputs was evaluated using the four parameters proposed by Shah et al. (2003): quantity, variety, novelty, and quality. Based on these metrics, creativity was compared across human outputs and generative AI ones. Subsequently, creativity was also analyzed across the different generative AI tools used in the experiment (see section "Preparation of the experiment").

By answering these research questions, this work seeks to contribute to the understanding of the current capabilities and limitations of image-generative AI systems in conceptual engineering design and to provide insights for the development of more effective human–AI design workflows. The contribution of this study lies in three main aspects. First, it provides a controlled comparison between human-generated and AI-generated conceptual design outputs within the same engineering design task. Second, it compares multiple image-generative AI systems rather than focusing on a single platform or model family. Third, it applies established engineering design creativity metrics (Quantity, Variety, Novelty, and Quality) to systematically evaluate visual concept generation by image-generative AI systems. In doing so, the study contributes to the ongoing discussion regarding the role of generative AI as an ideation support tool in conceptual engineering design.

Methodology

Preparation of the experiment

In order to address the research questions, an experiment was planned in two phases. In the first phase, design students would solve a given creative product design problem. In the second phase, AI image generation programs would be asked to solve the same statement. Subsequently, different creativity parameters would be analyzed and compared, with the aim of discerning the specific areas where AI demonstrates the greatest potential to assist designers.

The experiment enlisted the participation of 20 volunteers to form a convenience sample: 10 female and 10 male, with an average age of 23.2 years (standard deviation: 1.54). All participants were final-year students pursuing the Engineering Design degree. This research adhered to the American Psychological Association Code of Ethics and received approval from the Institutional Review Board at the Universitat Jaume I (CEISH/40/2022). Each participant provided informed consent. Design engineering students were selected because the study focuses specifically on conceptual engineering design tasks. Participants therefore required a minimum level of design knowledge. On the other hand, professional designers may exhibit more advanced expertise and potentially different creative strategies, so the use of students enabled the establishment of a consistent baseline appropriate for this exploratory study.

The selected statement for the experiment involved generating the maximum number of ideas that address the following problem: "Develop a new concept of a lunch box for university use." This specific problem was chosen because it appears in literature in studies of a similar nature (Chulvi et al., 2022). The solutions should be graphically represented, with each proposed alternative on separate sheets of paper. The allotted time for idea generation was set at 20 minutes, in line with the findings of Howard et al. (2009), which suggest that idea generation capacity significantly declines after 20 minutes. Additionally, this time constraint has been minimized, considering that the idea generation time for AI tools will always be shorter.

To select drawing AI tools with sufficient potential for creating creative product designs, a preliminary analysis of 43 different image generation AIs was conducted. For screening purposes, the following key characteristics were considered:

- The ability to generate distinct images even when the statement is repeated.
- Capability to present images in a rendering style similar to the drawings that a design engineering student might produce.
- Accessibility features.
- No additional cost with no loss of technical capabilities. The capacity to generate generic designs.
- Expert opinions regarding the technology they employ.

Ultimately, the chosen AI-based image-generation tools, labeled AI1-AI6 for convenience, were as follows (Table 1):

It is important to distinguish between the tool/platform level and the underlying model level. While six publicly accessible image-generation tools were selected for the study, these tools do not necessarily represent six entirely independent generative AI architectures. Some platforms operate using related diffusion-model families or derivative implementations. Therefore, the comparison conducted in this study primarily concerns the creative performance of accessible image-generative AI tools within realistic design-use scenarios rather than a strict benchmark comparison between fully independent foundational models.

All the AIs listed above are text-to-image generation tools, (image-generative AIs.) Their primary function is to create images

Table 1. AI tools used in the experience

	Tool	Underlying model family	Main feature	Website
AI1	Bing image creator	DALL·E 3	OpenAI proprietary model	https://www.bing.com/create
AI2	Leonardo AI	Stable Diffusion derivatives + proprietary finetunes	Artistic optimization	https://leonardo.ai/
AI3	Playground AI	SDXL	Diffusion-based	https://playgroundai.com/
AI4	Recraft AI	Proprietary Diffusion-like architecture	Vector-oriented features	https://recraft.ai/
AI5	SeaArt	Stable Diffusion derivatives	Anime/artistic specialization	https://www.seaart.ai/
AI6	Stablecog	Stable Diffusion	Open-source diffusion-based	https://stablecog.com/

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

based on textual descriptions as prompts. These tools are accessible online, making them available to a wide non-technical audience (Turchi et al., 2023). The platforms where these tools are available are user-friendly, with intuitive interfaces that allow users to quickly learn and use them. No coding or model training is required, as the underlying machine learning models are already trained on vast datasets of images and text. This enables the models to understand the meaning of a prompt and generate corresponding images.

While all these tools share a common goal, they differ in the specific models and techniques they employ.

- *DALL-E 3* excels at generating highly detailed and realistic images.
- *Leonardo AI* specializes in artistic styles and techniques, allowing users to create images in various artistic styles.
- *SeaArt AI* is focused on artistic expression, generating images with a strong artistic flair.
- *Playground AI* offers a user-friendly interface and a wide range of styles, making it accessible to users of all levels.
- *Recraft AI* provides tools for editing and refining generated images.
- *Stable Diffusion* is known for its versatility, capable of generating images in a wide range of styles, from realistic to abstract.

For the purposes of this study, only the core text-to-image generation capability of each tool was considered. Although some platforms may provide additional functionalities, such as image editing, video generation, or advanced paid features, these options were not used in the experiment and were not part of the selection criteria. Therefore, the comparison focuses exclusively on their ability to generate visual product concepts from textual prompts under default settings. The accessibility criterion referred to the possibility of using the text-to-image generation function without requiring paid features that would modify or enhance the core output capability. In this sense, "free access" was not considered as a technical characteristic of the AI model, but as a practical requirement to ensure that the selected tools could be used under comparable and reproducible conditions.

Development of the experiment

The experiment was conducted in two planned phases (Fig. 1). The first phase involved Design Engineering students, while the second phase artificial intelligence (AI) was involved.

In the initial experimental phase, voluntary participants were assembled and asked to sign the provided informed consent form. Each participant was then given drawing materials and sufficient sheets of paper. They were presented with a statement: "Develop a new concept of a lunch box to bring lunch to the university. Provide as many ideas as you can. Please draw each proposal on a separate sheet of paper."

In the second phase, AI outputs were collected. Since the number of generated ideas may influence creativity metrics such as Variety, Novelty, and Quality, output quantity was controlled at the level of each AI tool rather than at the level of total aggregated outputs. The purpose was not to make the total number of human and AI-generated concepts identical, but to evaluate each AI tool under two human-referenced output conditions.

Two reference points were selected based on human performance:

- Average Solutions ($\overline{X}_{sol.}$): The mean number of ideas generated by all human participants.
- Maximum Solutions ($Max_{sol.}$): The highest number of ideas generated by any individual participant.

These two values represent distinct but complementary human-referenced output conditions. The average reflects a typical human performance baseline, while the maximum represents the upper bound of human ideation capacity under the given conditions. These values were used as constraints in two rounds of AI prompts, each using the same design task:

- Round 1: AI tools were prompted to generate the same number of solutions as the average human participant (coded as $[\overline{X}_{sol.}]S\text{-}AI_n$).
- Round 2: AI tools were prompted to generate the maximum number of solutions achieved by human participants (coded as $[Max_{sol.}]S\text{-}AI_n$).

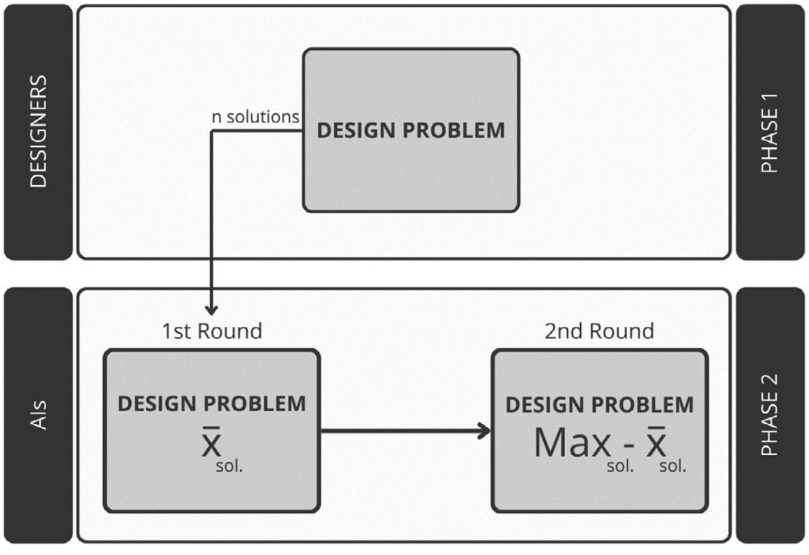


Figure 1. Methodology process.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

By using both values, the study captures not only standard performance levels but also peak ideation scenarios, enabling a more comprehensive comparison between human and AI creativity. The use of two quantity conditions allows us to assess whether the relative performance of AI compared to humans is sensitive to the number of generated ideas. This is particularly relevant given that AI systems can generate outputs at a much higher rate than humans, potentially biasing creativity assessments if output quantity is not controlled. However, the comparison should be interpreted at the level of participants/tools and human-referenced output conditions, rather than as a comparison based on identical total numbers of generated ideas.

Thus, the prompts used in the different AIs were, the first one "Develop a new concept of a lunch box to bring lunch to the university," and subsequently to $\overline{X}_{sol}$ and to $Max_{sol.}$ "Develop another different new concept of a lunch box to bring lunch to the university." Default models and default styles were used for the image generation. No generated images were classified as "failed generation," so all resulting outputs were included in the study.

Assessment method

To analyze the creativity resulting from the design process, the metric proposed by Shah et al. (2003) was employed. This metric distinguishes four parameters: quantity, variety, novelty, and quality. The definitions of these parameters are listed below:

Quantity: Refers to the number of solutions generated by the designer or design team. It is calculated by directly counting the outputs provided by the participants.

Variety: Represents the degree of diversity or difference among the generated solutions. The calculation considers how each function is performed for each output, based on the level of abstraction at which differentiation occurs. To assess variety, an organizational tree (example of Function 1 of D14 in Fig. 2) is constructed:

- At the first level, ideas are differentiated based on the Physical Principles (PP) that satisfy the same function. PP refers to the fundamental idea or concept underlying a creative solution. It is the scientific or technical basis that justifies how the idea works. Example: In the context of a lunch box, the physical principle refers to the way food is separated, such as physically dividing it into distinct compartments.
- At the second level, differentiation occurs according to Working Principles (WP). WP is the practical application of the physical principle. It describes how the physical principle is implemented in a functional design or system. Example: In the lunch box, the working principle would be the specific design used to achieve physical separation, such as a single container with internal compartments or multiple containers combined into one system.
- The third level distinguishes ideas based on their Embodiment (E). It refers to the way the working principle translates into a tangible design. It is the physical representation of the creative concept. Example: In the lunch box, embodiment refers to how multiple containers are physically connected, such as through fixed integration or detachable, modular connections.
- Finally, at the fourth and last level, differences at a Detailed level (D) are considered. This level refers to the individual components that make up the design. Each part must fulfill a specific function within the system. Example In the lunch box, this includes the specific design of the connection mechanism, such as the type of locking or fitting system used to assemble and disassemble the containers.

Variety is quantified using the refinement proposed by Nelson et al. (2009), which establishes the following formula for calculating the final variety score:

$$V = [10(n_{PP} - 1) + 5(n_{WP} - 1) + 2(n_E - 1) + (n_D - 1)]/N_{max}$$

where (n) represents the number of branches at the corresponding level of abstraction, and (N_{max}) is the maximum number of solutions.

Novelty: Refers to the degree of originality or rarity of the generated solutions compared to existing or known solutions. It is calculated as the weighted sum of the importance of key product subfunctions, multiplied by a coefficient corresponding to the level of originality with which these subfunctions have been addressed. This coefficient takes on a value of 0 if the subfunction is not fulfilled, 3 if it has been resolved in a very common manner, 7 if it has been resolved in an infrequent manner, and 10 if it has been resolved in a completely novel way. An example is shown in Table 2.

Quality: Represents the degree of suitability or effectiveness of the solutions generated according to problem criteria or requirements. This parameter is calculated based on the same subfunctions used

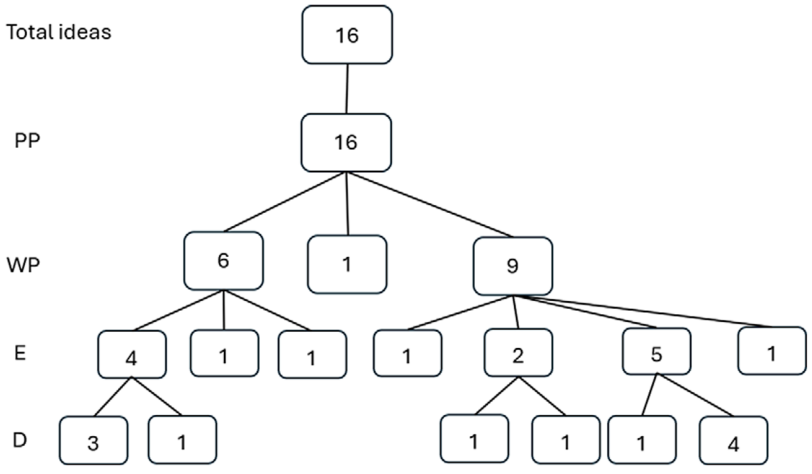


Figure 2. Example of an organizational tree used to evaluate variety (F1 of subject D14).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

Table 2. Example of novelty assessment

Level of novelty	0 – not fulfilled	3 – resolved in a very common manner	7 – resolved in an infrequent manner	10 – resolved in a completely novel way
Example of a fan – function: move the air	Does not move air	Standard fan, with blades and rotor	Uses centrifugal energy to optimize airflow	No blades, uses magnetism to move air

Table 3. Example of quality assessment

Level of quality	0	1	2	3	4	5	6	7	8	9	10
Example of a fan – function: easy to use	Inoperable					Basic and manual controls					Intelligent controls with connectivity to mobile applications

in assessing novelty. However, in this case, the effectiveness of the solution provided for each subfunction is evaluated on a scale from 0 to 10. Specifically, a value of 0 corresponds to a subfunction that is not fulfilled, a value of 5 indicates basic fulfillment, and a value of 10 indicates optimal resolution. The final quality score is also quantified through the weighted sum of the values considered for each subfunction, using the weight assigned to each subfunction. An example is shown in Table 3.

The same subfunctions and their weights considered by Chulvi et al. (2022) were also utilized for the design problem used as a task in the experimental procedure. In particular, the subfunctions considered were:

- Way of separating the food (weight: 0.6).
- Transport and/or storage after use (weight: 0.4).

All evaluations were conducted by an expert in the field of Design Engineering, with 17 years of experience in assessing creative design. A second expert, with 10 years of experience, realized independently the same evaluations in order to analyze the agreement between evaluators. The reliability of the agreement was calculated with a Fleiss Kappa assessment, with a value of 0.952. The value has also been recalculated without considering the values of the variable "quantity," since these values will always be coincident. In this case, the value of the Fleiss Kappa assessment is 0.857.

The study should be interpreted as an exploratory comparative design rather than as a conventional between-group experimental study. Human participants constituted a population sample with natural inter-individual variability, whereas the AI condition represented a selected set of image-generative systems evaluated under predefined generation settings. Consequently, the comparison does not aim to establish population-level equivalence between humans and AI, but rather to explore differences in creative performance under controlled conceptual design conditions.

Results

RQ1: Difference between human designers and image-generating AI outputs

Table 4 shows the creativity assessment results based on metrics proposed by Shah et al. (2003), including Quantity, Variety, Novelty, and Quality. A brief description of each metric, along with box plots for a visual comparison of the results, is provided. To provide an additional exploratory perspective, the four creativity dimensions were normalized using min–max normalization and averaged into a composite exploratory creativity score. Given the absence of a standardized weighting framework for aggregating Shah et al. (2003) creativity metrics, this score should be interpreted as a complementary descriptive indicator rather than as a definitive measure of creativity.

In terms of quantity, participants generated between 2 and 16 solutions (median: 3.5, mean: 4.45). In the first round, AIs were constrained to generate 4 solutions (4S-AI). This constraint was loosened in the second round, allowing AIs to generate 16 solutions (16S-AI), with solutions generated in batches of 4 until the target number was reached. It should be noted that Quantity was considered primarily as a process-related variable associated with the experimental generation settings rather than as a direct indicator of creative capability. Consequently, Quantity was not subjected to inferential statistical analysis nor interpreted independently.

In terms of Variety, Figure 3 shows that human designers consistently outperformed AIs, with participants achieving a median variety score of 0.45 and a mean of 0.60, compared to the much lower range of 0.08–0.17 for 4S-AI. When the number of solutions was increased to 16 (16S-AI), Variety scores improved slightly for some AIs, with AI6 achieving the highest score of 0.49, marginally exceeding the participants' median value but falling short of their mean.

Novelty scores, visualized in Figure 4 were more comparable between human designers and AI tools. In the 4S-AI condition, AIs achieved a median value equal to the students' median of 5.40, closely approximating the students' mean of 5.42. However, when generating 16 solutions, half of the AIs improved their novelty scores, approaching the maximum value of 7 achieved by students. This improvement shows that AI1, AI2, and AI5 can explore a broader solution space when allowed to generate more outputs. Nevertheless, the consistent novelty scores observed in AI3, AI4, and AI6 indicate that increasing the quantity of solutions does not always translate into higher novelty.

To assess quality, all solutions were evaluated. The maximum quality score achieved by each participant and AI was then identified for comparison. This approach was adopted because conceptual design processes involve an initial divergent ideation phase followed by a convergent concept-selection stage, in which the most promising alternative is identified for further development (Zheng et al., 2018). Consequently, the analysis focused on the highest-quality solution generated by each participant or AI system, as this best reflects their potential to produce viable concepts suitable for subsequent design stages. It should be noted that the maximum Quality score was calculated separately for each participant and each AI tool, rather than as a single maximum value for the human or AI group. Therefore, the comparison does not represent the best human solution versus the best AI solution, but the best solution

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

Table 4. Creativity assessment results

		Designer (D) / AI	Quantity	Variety	Novelty (max)	Quality (max)	Normalized composite creativity
First phase		D1	3	0.48	4.6	5.6	3.80
		D2	3	0.59	5.4	4.4	3.81
		D3	3	0.23	7	6.6	5.27
		D4	6	1.21	4	4	4.64
		D5	2	0.33	3	4	1.39
		D6	3	0.12	4.6	6.2	3.36
		D7	2	0.33	5.4	3.8	2.70
		D8	3	0.24	5.4	3.8	2.68
		D9	3	0.35	4.6	6.2	3.87
		D10	4	0.41	7	5.8	5.37
		D11	5	0.36	7.2	5.4	5.32
		D12	4	0.93	5.4	5.4	5.33
		D13	3	0.33	5.4	3.8	2.87
		D14	16	1.09	5.4	6	8.19
		D15	3	0.36	7	5.4	4.85
		D16	6	1.11	4.6	6.8	6.45
		D17	5	0.75	4.6	5.4	4.64
		D18	5	0.85	5.4	6.2	5.81
		D19	6	1.04	7	5	6.65
		D20	4	0.80	5.4	6.2	5.52
Second phase	First round	4S-AI1	4	0.13	5.4	5	3.32
		4S-AI2	4	0.08	5.4	3.8	2.50
		4S-AI3	4	0.13	5.4	5	3.32
		4S-AI4	4	0.12	5.4	3.8	2.59
		4S-AI5	4	0.15	5.4	3.4	2.42
		4S-AI6	4	0.17	5.4	2.6	1.98
	Second round	16S-AI1	16	0.40	7	5	7.02
		16S-AI2	16	0.37	7	5	6.95
		16S-AI3	16	0.19	5.4	5	5.60
		16S-AI4	16	0.33	5.4	3.8	5.20
		16S-AI5	16	0.17	7	5	6.51
		16S-AI6	16	0.49	5.4	3.8	5.55

generated by each individual designer or AI tool under the corresponding generation condition.

The analysis of Quality, shown in Figure 5, revealed that participants achieved higher scores overall, with a median of 5.4 and a mean of 5.3. The best-performing 4S-AI tools reached a maximum quality score of 5, which corresponds to the minimum threshold for functional proposals. Increasing the number of solutions to 16 led to modest improvements in Quality for some AIs, such as AI2, AI5, and AI6, while others maintained consistent or declining scores. Notably, two AIs failed to surpass the participants' minimum quality score of 3.8, even with the increased output. This suggests that AIs often struggle to produce high-quality solutions.

To validate these findings, further analyses were conducted using IBM SPSS Statistics (version 29.0.1.0) (Table 5). Preliminary normality tests (Shapiro–Wilk) revealed a non-normal distribution for Variety and Novelty ($p < .001$), while Quality and normalized composite creativity exhibited a normal distribution.

Consequently, non-parametric Mann–Whitney U tests were employed to compare student and AI performance on Variety and Novelty. Statistically significant differences were observed in Variety ($p = .002$). This indicates a substantial difference in the diversity in their outputs. However, no significant difference was found in Novelty ($p = .202$), suggesting that both students and AI generate ideas with comparable levels of novelty.

For Quality, which met the assumptions of normality, a parametric t-test was conducted. No significant difference was observed between students and AIs performance on this metric ($p = .573$). The normalized composite creativity analysis also revealed no

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

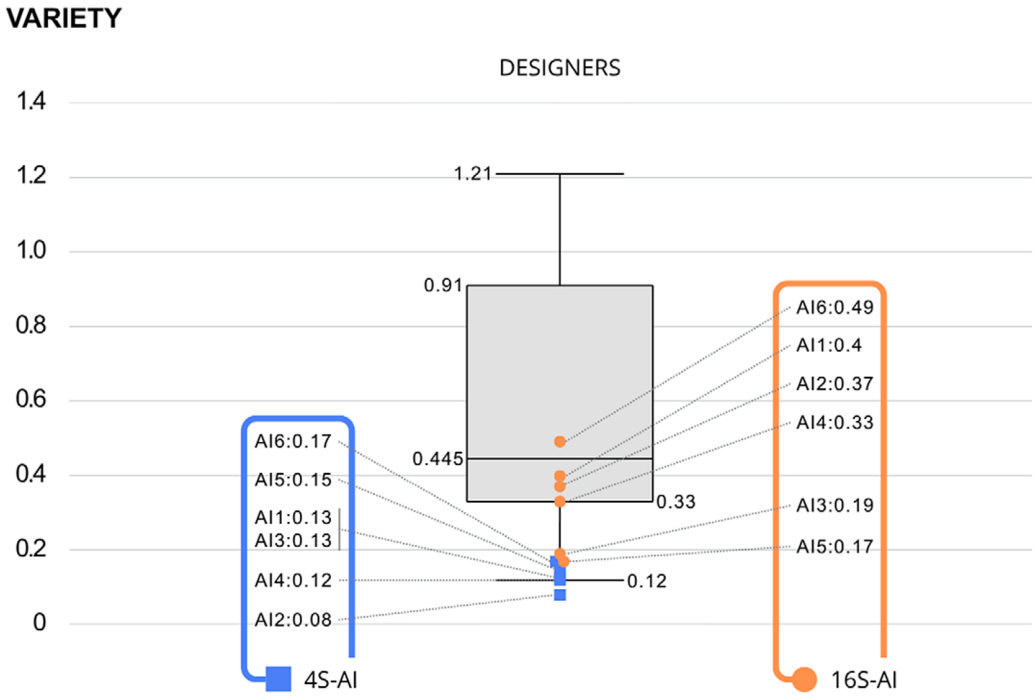


Figure 3. Box plot of variety results distribution.

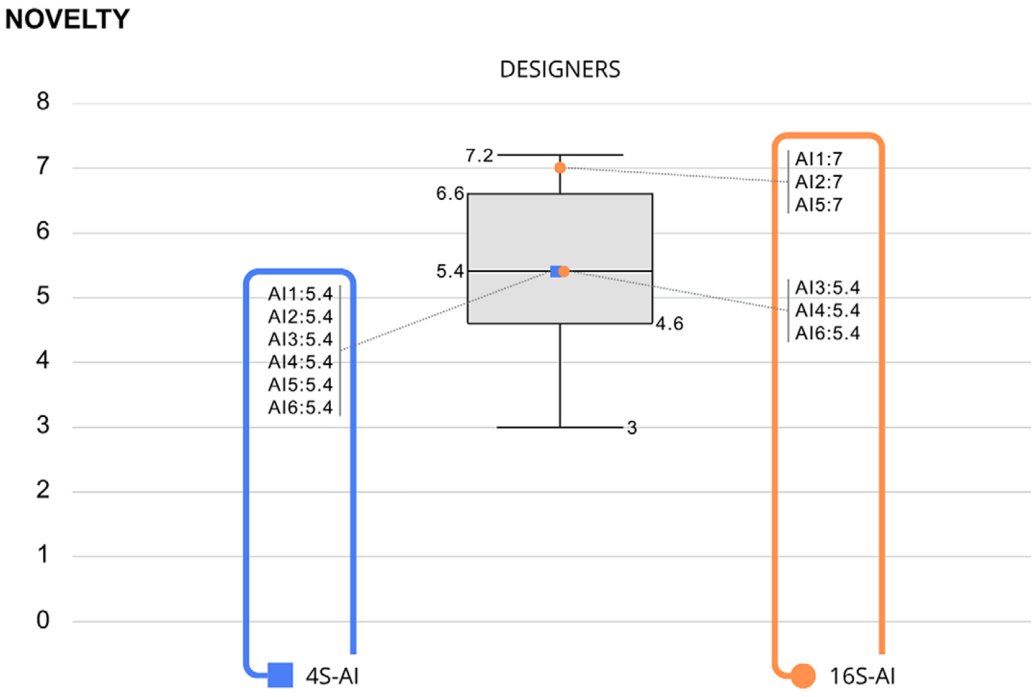


Figure 4. Box plot of novelty results distribution.

statistically significant differences between human designers and AI systems ($p = .738$).

These findings underscore the distinct creative capabilities of human designers, who leverage cognitive flexibility and interpretative skills to produce diverse and high-quality solutions.

RQ2: Difference between image-generative AIs outputs

Further comparisons among AI tools (Figs. 6–8) revealed variability in performance across metrics. The results compare the AIs under two conditions: when limited to generating 4 solutions (blue bars) and when allowed to generate up to 16 solutions (orange bars).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

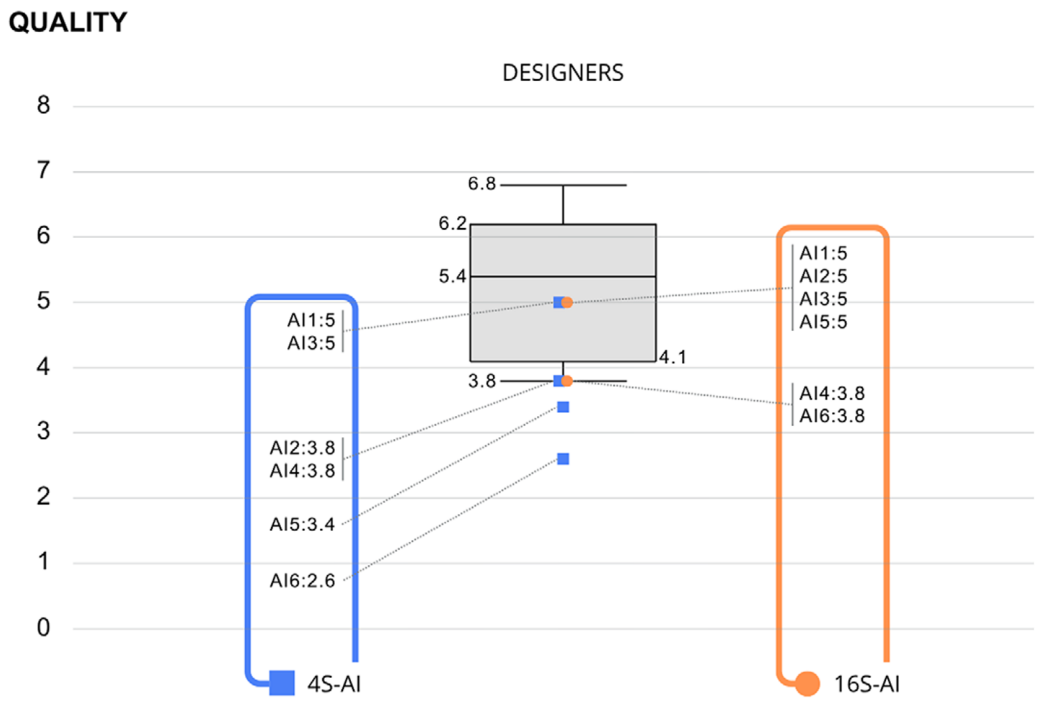


Figure 5. Box plot of quality results distribution.

Table 5. Results of Normality Shapiro–Wilk test and (non)parametric tests assessing the difference among participants and designers

Test	Normalized composite creativity	Variety	Novelty	Quality
Normality test Shapiro–Wilk	$p = .223$	$p < .001$	$p < .001$	$p = .125$
Mann–Whitney *U* tests (non-parametric)		$p = .002$	$p = .202$	–
T-test (parametric)	$p = .738$	–	–	$p = .573$

Comparisons among individual AI systems were treated descriptively, and paired statistical analyses were also conducted comparing the 4-solution (4S) and 16-solution (16S) generation settings across the six AI tools. Due to the limited sample size and the non-normal distribution of several variables, Wilcoxon signed-rank tests were employed.

Figure 6 shows that asking AI tools to generate a larger number of solutions (16 in this case) can lead to an increase in the variety of the generated outputs. However, the extent of this increase varies across different AI tools.

Overall, the AI tools generally achieved higher variety scores when asked to generate 16 solutions compared to when limited to

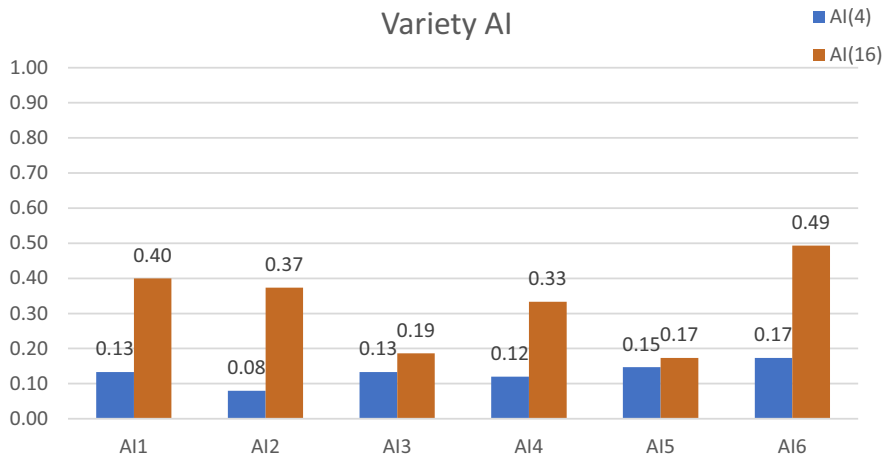


Figure 6. Comparison of AI outputs in terms of variety.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

4 solutions. The results of the statistical analysis revealed a significant increase in Variety when AI systems generated 16 solutions instead of 4 (W = 0.0, p = .031), suggesting that expanding the number of generated outputs improves the exploration of the solution space. From the descriptive analysis, it can be perceived that some tools show a more substantial increase in variety score (e.g., AI1, AI2, AI4, AI6), while others show a smaller increase (e.g., AI3, AI5). AI tool AI6 stands out with the highest variety score among all tools, especially when generating 16 solutions. This suggests that AI6 is particularly effective at producing a wide range of diverse solutions.

Figure 7 shows that asking an AI tool to generate a larger number of solutions can lead to an increase in the novelty of the generated outputs in some of the tools tested. However, these differences did not reach statistical significance ($W = 0.0$, $p = .083$).

While some tools exhibit increasing novelty with more solutions (AI1, AI2, AI5), others maintain a consistent score (AI3, AI4, AI6). In particular:

- Increased Novelty: AI1, AI2, and AI5 show an increase in novelty scores when generating a larger number of outputs. This suggests that these AI tools can generate more novel solutions when allowed to explore a wider solution space.
- Consistent Novelty: AI3, AI4, and AI6 exhibit consistent novelty scores regardless of whether 4 or 16 solutions are generated. This result suggests that for some AI tools, increasing the number of solutions generated may not always translate to a significant improvement in the novelty of the outputs.

Figure 8 shows that the relationship between the number of solutions generated by AI tools and the quality of those solutions is not straightforward. Some tools maintain high quality with more solutions, while others experience a decline in quality as the number of solutions increases. Consequently, the statistical analysis shows no significant difference in case of quality ($W = 0.0$, $p = .109$).

The quality of AI outputs varies significantly depending on the AI tool and the number of solutions generated.

- Consistent Quality: AI1, AI3, and AI4 exhibit consistent output quality regardless of whether 4 or 16 solutions are generated. Increasing the number of outputs does not noticeably improve the quality in these cases.
- Improved Quality with Increased Outputs: AI2, AI5, and AI6 demonstrate an increase in output quality when the number of generated solutions is increased from 4 to 16. AI5 shows the most pronounced improvement in this regard. However, despite the

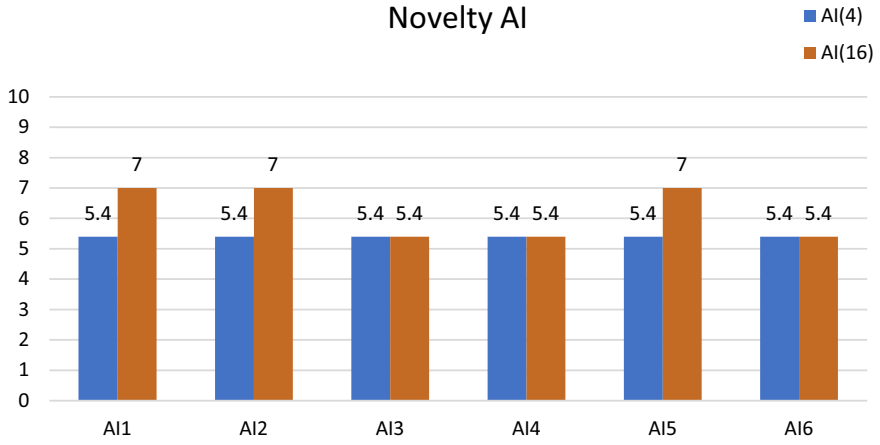


Figure 7. Comparison of AI outputs in terms of Novelty (the maximum value is considered).

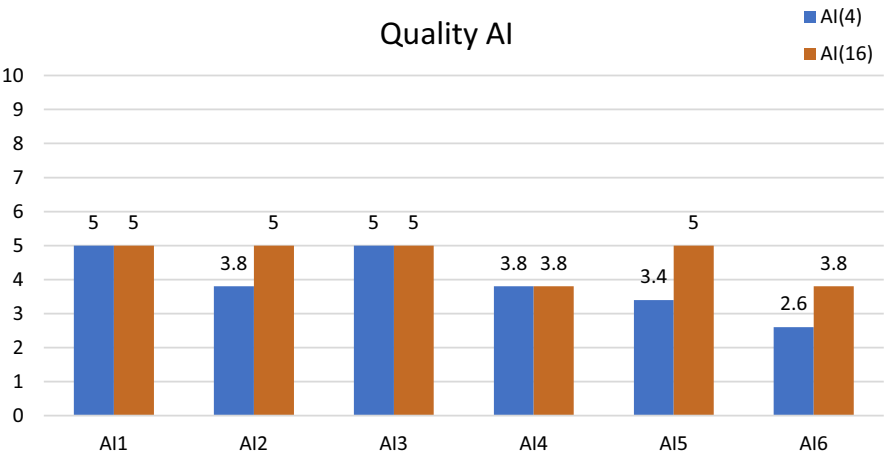


Figure 8. Comparison of Ai outputs in terms of Quality (maximum values are considered).

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

increase in the number of outputs, AI6 displays the lowest quality scores, both when generating 4 and 16 solutions.

These findings suggest that while increasing the number of generated solutions can potentially enhance the overall quality of AI outputs, this effect is highly dependent on the specific AI tool employed.

Overall, some tools, such as AI6, excelled in Variety when generating 16 solutions, while others, like AI5, demonstrated notable improvements in Quality. However, these improvements often failed to surpass the benchmarks set by human designers.

Discussion

RQ1: Differences between human designers and image-generative AI outputs

This section discusses the comparative analysis of creative outputs from human designers and AI image generators, focusing on Quantity, Variety, Novelty, and Quality. The findings reveal distinct strengths and weaknesses for each, highlighting their respective creative processes.

- *Quantity*: As expected, AI tools demonstrated a clear advantage in generating large numbers of solutions within very short timeframes due to their computational processing capacity. Human designers produced between 2 and 16 solutions (median = 3.5). However, the study did not aim to compare raw production speed between humans and AI systems, as such a comparison would primarily reflect differences in computational efficiency rather than creative performance. Instead, the comparison was designed around a controlled output-based fairness principle, in which AI systems were requested to generate numbers of solutions comparable to those observed in the human sample. This approach was adopted to enable a more balanced comparison of the creative characteristics of the generated concepts, particularly Variety, Novelty, and Quality, rather than differences in production rate. Consequently, Quantity was interpreted primarily as an experimental process variable rather than as a standalone indicator of creative capability.
- *Variety*: significantly differs across human designers who outperformed AIs. The median variety score for humans was 0.45 (mean 0.60), while 4S-AIs scored significantly lower, ranging from 0.08 to 0.17. While increasing the number of AI-generated solutions to 16 (16S-AI) led to some improvement (e.g., AI6 reaching 0.49), no AI surpassed the human mean. This highlights a key limitation of current AI tools: a high degree of fixation, resulting in structurally or conceptually similar outputs. Even the highest-performing AI in the 16S condition only marginally exceeded the median human score, falling short of the mean. This reinforces the observation that AIs struggle to produce truly diverse solutions, and simply generating more outputs only partially mitigates this issue while simultaneously increasing the burden of evaluation for the user. The significant difference in variety (p = .002) further underscores this point.
- *Novelty*: In contrast to Variety, AI performance in Novelty was comparable to human designers. The median novelty score for 4S-AIs (5.4) matched the human median, and the mean (5.42) was also similar. Interestingly, when generating 16 solutions, some AIs (AI1, AI2, AI5) achieved higher novelty scores, reaching the maximum value observed among human designers (7). However, other AIs (AI3, AI4, AI6) showed no improvement with increased output. This suggests that while some AI tools can explore a wider solution space when generating more outputs, others are constrained by their underlying algorithms. The consistent novelty scores among 4S-AIs might indicate a shared underlying logic in their creative processes, though further investigation is needed. While some AIs demonstrated the potential for high novelty, the statistical analysis (p = .202) showed no significant difference between humans and AIs in this metric.
- *Quality*: based on the descriptive statistics results, human designers consistently achieved higher Quality scores (median 5.4, mean 5.3). In the 4S-AI condition, the highest-scoring AIs reached a quality score of 5, the minimum threshold for functional proposals. Increasing the number of solutions to 16 led to modest improvements for some AIs (AI2, AI5, AI6), but two AIs failed to surpass the minimum quality score achieved by human students (3.8), even with increased output. This suggests that while AIs can produce functional outputs, they struggle to achieve the nuanced, high-quality solutions characteristic of human design. AIs often remain at the minimum threshold of required quality, potentially due to their literal interpretation of the problem statement. This highlights the importance of human interpretation and reformulation of design briefs to effectively guide AI tools. Despite the observed differences, the statistical analysis did not reveal a significant difference in quality (p = .573).

Overall, the results highlight the complementary roles of human designers and AI in creative processes. Although AIs are effective at generating large numbers of solutions quickly, their outputs often lack the diversity, novelty, and nuanced quality demonstrated by human designers. These findings suggest that AI systems are currently more effective as support tools for ideation rather than as autonomous substitutes for human creativity. Consequently, human oversight and interpretative guidance remain essential to direct AI-generated outputs toward more meaningful, contextually appropriate, and innovative design solutions. Furthermore, since the study involved design engineering students rather than professional designers, the findings should be interpreted cautiously when considering professional design practice. Experienced designers may demonstrate more advanced ideation strategies and evaluation capabilities, potentially increasing the differences observed between human and AI creative performance.

RQ2: Differences among image-generative AI outputs

The comparison of AI tools under two conditions (4S-AI and 16S-AI) revealed variability in performance across the creativity metrics, underscoring the influence of model architecture and generation parameters.

- *Variety*: Increasing the number of solutions from 4 to 16 significantly led to higher Variety scores. AI6 (Stablecog) achieved the highest Variety score (0.49) when generating 16 solutions, indicating its ability to produce a more diverse range of outputs. The extent of improvement in Variety scores varied significantly among tools. For example, AI3 (Playground AI) and AI5 (SeaArt) showed minimal gains, suggesting that some tools may have inherent limitations in their ability to diversify outputs, even when given more opportunities to do so.

  These findings highlight the importance of system-specific adjustments to enhance diversity, as certain architectures or algorithms may exhibit a higher tendency toward fixation or repetitive outputs.
- *Novelty*: The relationship between the number of solutions and Novelty scores varied across tools, although the differences observed were not significant. AI1 (DALL·E 3), AI2 (Leonardo AI),

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

and AI5 (SeaArt) exhibited increased Novelty scores when generating 16 solutions. This suggests that these tools can explore a broader solution space and produce more original outputs when given more opportunities to iterate. In contrast, AI3 (Playground AI), AI4 (Recraft AI), and AI6 (Stablecog) maintained consistent Novelty scores regardless of the number of outputs generated. This indicates that for some tools, increasing the number of outputs does not necessarily lead to enhanced originality, possibly due to limitations in their generative algorithms or reliance on predefined patterns. While some AI tools benefit from generating more solutions in terms of originality, others may require algorithmic refinements to break free from inherent constraints.

- *Quality*: The impact of increasing output quantity on Quality was inconsistent and not statistically significant. However, AI tools such as AI2 (Leonardo AI), AI5 (SeaArt), and AI6 (Stablecog) demonstrated improvements in Quality when generating 16 solutions. This suggests that, for these tools, exploring a larger solution space may help identify higher-quality outputs. Tools like AI1 (DALL·E 3), AI3 (Playground AI), and AI4 (Recraft AI), maintained consistent Quality scores regardless of the number of outputs. This indicates that their performance is relatively stable and less influenced by the quantity of solutions generated.

Despite excelling in Variety, AI6 (Stablecog) consistently achieved the lowest Quality scores. This suggests a potential trade-off between diversity and quality, where efforts to maximize diversity may lead to outputs that lack refinement or practical value. The varying effects of increased output quantity on Quality underscore the need to balance diversity and quality in AI-generated solutions. Tailoring the parameters or post-processing techniques for each tool could help optimize outputs across multiple metrics.

The results underscore the complexity of optimizing AI performance across creativity metrics. While increasing the number of generated solutions generally enhances Variety, it has not been demonstrated with statistical significance Novelty and Quality are improved. The findings are consistent with recent research suggesting that generative AI systems tend to support divergent exploration more effectively than convergent refinement or final concept evaluation (Wang et al., 2025c). The observed variability across AI tools suggests that differences in generation pipelines, interface implementations, prompting behavior, and underlying model families may influence creative performance outcomes. For example, AI6 (Stablecog) demonstrates high performance in Variety while showing comparatively lower Quality scores, indicating potential trade-offs between exploratory diversity and output refinement. These findings reinforce the importance of human intervention and supervision to achieve balanced and contextually appropriate design outcomes.

The study should therefore be interpreted as an evaluation of practically accessible image-generative design tools rather than a strict architectural comparison between independent foundational AI models.

Conclusion

Creativity is a cornerstone of design engineering, fueling innovation and solving complex problems. Visual representations play a pivotal role in this process, serving as essential tools for communicating and refining design ideas. The emergence of image-generative AI offers new opportunities to augment these creative processes, yet the extent of their capabilities, particularly in comparison to human designers, remains an area of active exploration. This exploratory study aims to evaluate the creativity of several AI tools and compare their outputs with those of human designers using the creativity metrics of Quantity, Variety, Novelty, and Quality, as defined by Shah et al. (2003). A sample of convenience of 20 design engineering students and six image-generative AI systems were tasked with the same design challenge, with humans constrained by time and AIs limited by predefined output parameters.

The answer to RQ1 regarding the comparison between human creativity and AI creativity follows; the findings reveal some differences in the creative capacities of humans and AIs. While AIs excel in generating large quantities of solutions quickly, this strength reflects computational efficiency rather than true creative ability. Human designers far outperformed AIs in terms of Variety, with the AI outputs exhibiting a pronounced fixation and limited conceptual diversity. Even when generating 16 solutions, only two AIs (Bing Image Creator and Stablecog) approached the median Variety score of human designers, and none achieved the average. This suggests persistent limitations in the evaluated AI tools' ability to consistently generate diverse and conceptually distinct outputs within the context of conceptual engineering design tasks. In case of Quality, humans performed overall better than AI, but this difference was not statistically significant.

The results related to the comparison of creativity among AI systems (RQ2), reveal the complexities of AI performance across various creativity metrics. While increasing the number of generated solutions generally improves Variety and, in some cases, Novelty, it does not consistently enhance Quality. The observed variability across the evaluated AI tools suggests that differences in generation pipelines, prompt interpretation mechanisms, implementation strategies, and underlying model families may influence creative performance outcomes. For instance, tools like AI6 (Stablecog) demonstrated strong performance in Variety while showing comparatively lower Quality scores, indicating possible trade-offs between exploratory diversity and output refinement. This highlights the need for human intervention and supervision to achieve balanced outputs. Regarding Novelty, AI tools showed more promise, with most achieving scores comparable to the human median, and some (e.g., Bing Image Creator, Leonardo AI, and SeaArt) even surpassing it when generating 16 solutions, indicating their capacity to explore broader solution spaces, although some outputs still exhibited constrained conceptual exploration. Quality, however, consistently posed a challenge.

Although some tools (e.g., Bing Image Creator and Playground) approached the human median when generating four solutions, no AI surpassed the human average, even with increased output. This emphasizes the critical role of human interpretation and refinement in achieving high-quality results, as AIs often struggle to fully capture the nuances of design tasks. Despite these limitations, the AI tools tested show potential as valuable aids in the early stages of the creative process. While often lacking in Variety and Quality, AI-generated outputs can provide a foundation of preliminary ideas to inspire human designers. Their relatively strong performance in Novelty suggests they can introduce fresh perspectives, even if these require refinement. Therefore, designers must actively guide AI systems through well thought prompts and variable configuration to maximize the relevance and utility of AI-generated solutions, positioning AI as a complementary tool – or a "partner" (Lubart, 2025) – rather than a replacement for human creativity. However, since several evaluated tools rely on related diffusion-based model

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

families, these findings should not be interpreted as a strict comparison between fully independent foundational AI architectures.

Limitations

Several contextual factors should be considered when interpreting these results. Firstly, the study relied on a convenience sample composed of design engineering students rather than professional designers. Although this population is commonly used in creativity and engineering design research, the relatively homogeneous educational background and experience level of the participants limit the generalizability of the findings. Consequently, the results should not be directly extrapolated to professional design practice without caution. It is reasonable to assume that experienced designers would demonstrate more advanced ideation strategies, domain knowledge, and concept-evaluation capabilities, potentially widening the observed differences between humans and AI systems (Bonnardel and Marmèche, 2004). This aligns with findings in other domains, such as the study by Noy and Zhang (2023), which showed that lower-ability participants benefited more substantially from AI assistance than higher-performing individuals.

Secondly, the students generated ideas without the support of a structured design methodology. Prior research suggests that systematic ideation methods can significantly influence creative outcomes and design-space exploration (Chulvi et al., 2013). Therefore, the absence of a formal design methodology may have affected the creative performance observed in both human and AI-assisted conditions.

Finally, the findings should also be interpreted considering the rapidly evolving nature of image-generative AI technologies. The evaluated tools represent specific versions and implementations available at the time of the experiment, and many of these systems undergo frequent updates affecting their underlying models, default parameters, prompting behavior, interface design, and generation capabilities. Consequently, the results should not be interpreted as stable or definitive assessments of the long-term creative potential of these AI systems, but rather as a snapshot of their performance under the technological conditions existing during the study period. Furthermore, the statistical inferences derived from the study should be interpreted cautiously and within the boundaries of the evaluated sample and AI configurations. The findings are intended to provide exploratory insights into current image-generative AI capabilities rather than definitive generalizations regarding human and artificial creativity. Future directions.

This study provides an initial foundation for developing a collaborative methodology between AI and designers to optimize the creative idea generation process. Future work should further analyze and contrast this potential joint effort. Several avenues for expansion exist, starting with the development of prompts that range from more to less constrained for AI systems, along with clear instructions for the design task assigned to human designers; this approach would allow for a comprehensive comparison of the results across all tests conducted in this experiment. Another direction could be investigating the combination of text-based AI with image-generative AI to further aid the design process, and exploring how integrating both with specific design methodologies impacts outcomes. The observed limitations of AIs, particularly in Variety and Quality, highlight areas for improvement in AI algorithms and model architectures. Integrating structured design methodologies could enhance the creative outputs of both humans and AIs, potentially narrowing performance gaps.

As AI technologies continue to evolve, their capabilities in creative tasks are expected to improve, but this study underscores the irreplaceable value of human expertise, particularly in interpreting and refining design briefs to produce high-quality outcomes. Future research should therefore focus on exploring the synergy between AI and professional designers, expanding on the methodologies developed here to optimize the creative process and establish AI as a powerful ally in advancing innovation and problem-solving within design engineering.

References

Agogué M, Kazakçi A, Hatchuel A, Le Masson P, Weil B, Poirel N and Cassotti M (2014) The impact of type of examples on originality: Explaining fixation and stimulation effects. *The Journal of Creative Behavior* **48**(1), 1–12. https://doi.org/10.1002/jocb.37.

Amabile TM (1983) The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology* **45**(2), 357–376. https://doi.org/10.1037/0022-3514.45.2.357.

Amabile TM (2019) GUIDEPOST: Creativity, artificial intelligence, and a world of surprises Guidepost letter for academy of management discoveries. *Academy of Management Discoveries* **6**(3), 351–354. https://doi.org/10.5465/amd.2019.0075.

Berni A, Maccioni L and Borgianni Y (2020) Observing pictures and videos of creative products: An eye tracking study. *Applied Sciences* **10**(4), 1480. https://doi.org/10.3390/app10041480.

Bonnardel N and Marmèche E (2004) Evocation processes by novice and expert designers: Towards stimulating analogical thinking. *Creativity and Innovation Management* **13**(3), 176–186. https://doi.org/10.1111/j.0963-1690.2004.00307.x.

Bordas A, Le Masson P, Thomas M and Weil e B (2024) What is generative in generative artificial intelligence? A design-based perspective. *In Design Computing and Cognition'24. Montréal* **35**(4), 427–443. https://doi.org/10.1007/s00163-024-00441-x.

Borgianni Y, Maccioni L, Fiorineschi L and Rotini F (2020) Forms of stimuli and their effects on idea generation in terms of creativity metrics and non-obviousness. **8**(3), 147–164. https://doi.org/10.1080/21650349.2020.1766379.

Bouschery SG, Blazevic V and Piller FT (2023) Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management* **40**(2), 139–153. https://doi.org/10.1111/jpim.12656.

Brisco R, Hay L and Dhami S (2023) Exploring the role of text-to-image AI in concept generation. *Proceedings of the Design Society* **3**, 1835–1844. https://doi.org/10.1017/pds.2023.184.

Cardoso C and Badke-Schaub P (2011) The influence of different pictorial representations during idea generation. *The Journal of Creative Behavior* **45**(2), 130–146. https://doi.org/10.1002/j.2162-6057.2011.tb01092.x.

Chen HC and Chen Z (2023) Using chatgpt and midjourney to generate chinese landscape painting of tang poem 'the difficult road to shu'. *International Journal of Social Sciences* **3**(2), 1–10. https://doi.org/10.35745/ijssai2023v03.02.0001.

Chen L, Song Y, Guo J, Sun L, Childs P and Yin Y (2025) How generative AI supports human in conceptual design. *Design Science* **11**, e9. https://doi.org/10.1017/dsj.2025.2.

Chulvi V (2025) Effectiveness of the AI using different typologies of design methods. *Research in Engineering Design* **36**(2), 1–12. https://doi.org/10.1007/s00163-025-00449-x.

Chulvi V, González-Cruz MC, Mulet E and Aguilar-Zambrano J (2013) Influence of the type of idea-generation method on the creativity of solutions. *Research in Engineering Design* **24**(1), 33–41. https://doi.org/10.1007/s00163-012-0134-0.

Chulvi V, Royo M, Agost MJ, García-García C and Ruiz-Pastor L (2022) Relaxing or arousing environments (through color and music) for fostering designer's creativity. *International Journal of Design Creativity and Innovation* **10**(4), 247–260. https://doi.org/10.1080/21650349.2022.2107081.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.

Chulvi V, **Ruiz-Pastor L**, **Royo M and García-García C** (2025) Design creativity in AI: Using the SCAMPER method. *AI EDAM* **39**, e13. https://doi.org/10.1017/S0890060425100036.

Cropley D (2023) Is artificial intelligence more creative than humans? *Learning Letters* **2**, 13–13. https://doi.org/10.59453/ll.v2.13.

Dai S, **Li Y**, **Grace K and Globa A** (2023) Towards human-AI collaborative architectural concept design via semantic AI. In *International Conference on Computer-Aided Architectural Design Futures*. Cham: Springer Nature Switzerland, pp. 68–82. https://doi.org/10.1007/978-3-031-37189-9_5.

Doshi AR and Hauser O (2023) Generative artificial intelligence enhances creativity but reduces the collective diversity of novel content. Available at SSRN. https://doi.org/10.2139/ssrn.4535536.

Epstein Z, **Hertzmann A**, **Akten M**, **Farid H**, **Fjeld J**, **Frank MR**, et al. (2023) Art and the science of generative AI. *Science (New York, N.Y.)* **380**(6650), 1110–1111. https://doi.org/10.1126/science.adh4451.

Filippi S (2023) Measuring the impact of ChatGPT on fostering concept generation in innovative product design. *Electronics* **12**(16), 3535. https://doi.org/10.3390/electronics12163535.

Fui-Hoon Nah F, **Zheng R**, **Cai J**, **Siau K and Chen L** (2023) Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research* **25**(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814.

Guilford JP (1950) Creativity. *American Psychologist* **5**(9), 444–454. https://doi.org/10.1037/h0063487.

Haefner N, **Wincent J**, **Parida V and Gassmann O** (2021) Artificial intelligence and innovation management: A review, framework, and research agenda. *Technological Forecasting and Social Change* **162**, 120392. https://doi.org/10.1016/j.techfore.2020.120392.

Howard TJ, **Culley SJ and Dekoninck EA** (2009) Stimulating creativity: A more practical alternative to TRIZ. In Bergendahl MN, Grimheden M, Leifer L, Skogstad P and Lindemann U (eds), *17th International Conference on Engineering Design*. Stanford, California: ICED'09.

Hutson J and Cotroneo P (2023) Generative AI tools in art education: Exploring prompt engineering and iterative processes for enhanced creativity. *Meta* **4**(1), 14. https://doi.org/10.54517/m.v4i1.2164.

Ishihara S, **Kuo R and Ishihara K** (2023) AI image generation boosts Kansei engineering design process. *Kansei Engineering* **101**(101). https://doi.org/10.54941/ahfe1002988.

Kılınç HK and Keçecioğlu ÖF (2024) Generative artificial intelligence: A historical and future perspective. *Academic Platform Journal of Engineering and Smart Systems* **12**(2), 47–58. https://doi.org/10.21541/apjess.1398155.

Lubart T (2005) How can computers be partners in the creative process: Classification and commentary on the special issue. *International Journal of Human-Computer Studies* **63**(4–5), 365–369. https://doi.org/10.1016/j.ijhcs.2005.04.002.

Mhlanga D (2024) Generative AI for emerging researchers: The promises, ethics, and risks. *Ethics, and Risks* (February 24, 2024). http://doi.org/10.2139/ssrn.4737492.

Nelson BA, **Wilson JO**, **Rosen D and Yen J** (2009) Refined metrics for measuring ideation effectiveness. *Design Studies* **30**(6), 737–743. https://doi.org/10.1016/j.destud.2009.07.002.

Noy S and Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. Available at SSRN 4375283. https://doi.org/10.2139/ssrn.4375283.

Oktradiksa A, **Bhakti CP**, **Kurniawan SJ and Rahman FA** (2021) Utilization artificial intelligence to improve creativity skills in society 5.0. *Journal of Physics: Conference Series* **1760**(1), 012032. https://doi.org/10.1088/1742-6596/1760/1/012032.

Paananen V, **Oppenlaender J and Visuri A** (2024) Using text-to-image generation for architectural design ideation. *International Journal of Architectural Computing* **22**(3), 458–474. https://doi.org/10.1177/14780771231222783.

Pei E, **Campbell I and Evans M** (2011) A taxonomic classification of visual design representations used by industrial designers and engineering designers. *The Design Journal* **14**(1), 64–91. https://doi.org/10.2752/175630610X12877385838803.

Ren L and Du M (2024, January) From canvas to code: The evolution of generative art in the AI era. In *ADDT 2023: Proceedings of the 2nd International Conference on Art Design and Digital Technology*, ADDT 2023, September 15–17, 2023, Xi'an, China (p. 479). European Alliance for Innovation. https://doi.org/10.4108/eai.15-9-2023.2340877.

Shah JJ, **Smith SM and 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.

Stock-Homburg R (2023) The tightrope between human and AI-generated innovation: A turing test. Available at SSRN 4610978. http://doi.org/10.2139/ssrn.4610978.

Turchi T, **Carta S**, **Ambrosini L and Malizia A** (2023) Human-AI co-creation: Evaluating the impact of large-scale text-to-image generative models on the creative process. In Spano LD, Schmidt A, Santoro C and Stumpf S (eds), *End-User Development. IS-EUD 2023. Lecture Notes in Computer Science*, Vol. 13917. Cham: Springer. https://doi.org/10.1007/978-3-031-34433-6_3.

Townsend C and Kahn BE (2014). The "visual preference heuristic": The influence of visual versus verbal depiction on assortment processing, perceived variety, and choice overload. *Journal of Consumer Research*, **40**(5), 993–1015. https://doi.org/10.1086/673521

Uludag K (2023) Testing creativity of ChatGPT in psychology: Interview with ChatGPT. Available at SSRN 4390872. http://doi.org/10.2139/ssrn.4390872.

Wang D and Han J (2023) Exploring the impact of generative stimuli on the creativity of designers in combinational design. *Proceedings of the Design Society* **3**, 1805–1814. https://doi.org/10.1017/pds.2023.181.

Wang B, **Han J**, **Zhao X**, **Yin Y**, **Chen L and Childs P** (2025a) Creative combinational design through generative AI in different dimensional representations: An exploration. *Design and Artificial Intelligence* **1**(1), 100006. https://doi.org/10.1016/j.daai.2025.100006.

Wang N, **Kim H**, **Peng J and Wang J** (2025b) Exploring creativity in human–AI co-creation: A comparative study across design experience. *Frontiers in Computer Science* **7**, 1672735. https://doi.org/10.3389/fcomp.2025.1672735.

Wang P, **Khinvasara Y**, **Creijghton GJ**, **Scholing T**, **Wang Y**, **Zhou Z**, **Childs P and Yin Y** (2025c) Enhancing designer creativity through human–AI co-ideation: A co-creation framework for design ideation with custom GPT. *AI EDAM* **39**, e22. https://doi.org/10.1017/S0890060425100127.

Wang B, **Zhu Y**, **Chen L**, **Liu J**, **Sun L and Childs P** (2023) A study of the evaluation metrics for generative images containing combinational creativity. *AI EDAM* **37**, e11. https://doi.org/10.1017/S0890060423000069.

Zhang H, **Chen P**, **Xie X**, **Jiang Z**, **Zhou Z and Sun L** (2024) A hybrid prototype method combining physical models and generative artificial intelligence to support creativity in conceptual design. *ACM Transactions on Computer-Human Interaction* **31**(5), 1–34. https://doi.org/10.1145/3689433.

Zheng X, **Ritter SC and Miller SR** (2018) How concept selection tools impact the development of creative ideas in engineering design education. *Journal of Mechanical Design* **140**(5), 052002. https://doi.org/10.1115/1.4039338.

Downloaded from https://www.cambridge.org/core. 15 Nov 2026 at 20:20:36, subject to the Cambridge Core terms of use.