Proceedings of the ASME 2022
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC-CIE2022
August 14-17, 2022, St. Louis, Missouri

DETC2022-87505

A STUDY OF THE EXPLORATORY CREATIVITY PERFORMANCE BETWEEN MACHINE AND HUMAN DESIGNERS

Yuan Yin
Imperial College London
London, United Kingdom

Kaitong Qin
Zhejiang University
Zhejiang, China

Huiting Liu
Zhejiang University
Zhejiang, China

Peter Childs
Imperial College London
London, United Kingdom

Lingyun Sun
Zhejiang University - China Southern Power Grid Joint
Research Centre on AI, Zhejiang University
Zhejiang, China

Liuqing Chen*
Zhejiang University
Zhejiang, China
(*corresponding author)

ABSTRACT

Exploratory creativity (E-creativity) is used to represent the creative performance behind the exploration process when establishing conceptual space. Researchers have attempted to build computational E-creativity models to help human generate more creative ideas or solutions. This trend sparks the discussion on whether the performance of machine can achieve a similar level to human beings. However, the performance gap of E-creativity between human beings and machine has not been fully studied. This study aims to investigate the E-creativity performance differences between machine and human designers. To be specific, a state-of-the-art model DALL·E is chosen as a representative of machines for generating E-creativity imagery and is compared to novice designers who are the representative for generating E-creativity imagery of humans. Expert designers are recruited as assessors to assess the creativity and E-creativity performance of the collected human and machine data. The experimental results reveal that the creativity level of humans is higher than that of machine. The E-creativity level of machine is higher than that of humans. The textual E-creativity performance is higher than the imagery E-creativity performance of humans while it is lower than the imagery E-creativity performance of machine. The results provide insights for supporting the development of more advanced E-creativity engines and corresponding evaluation methods.

Keywords: Exploratory creativity; creativity evaluation; computational creativity

1. INTRODUCTION.

Creativity is essential for design innovation [1]. Computational creativity is promoted due to the advance of computer science and machine learning technologies [2, 3]. Computational creativity consists of a set of mechanisms that can achieve or simulate creative behaviors [3]. When mentioning creativity, people tend to consider it as human creativity. The promotion of computational creativity triggers the discussion on whether the performance of machine can achieve a similar level to human beings [4].

As one kind of creativity, exploratory creativity (E-creativity) is used to represent the creative performance behind the exploration process when establishing conceptual space or style [5-7]. However, the performance gap of E-creativity between human beings and machine has not been fully researched. This study aims to investigate the E-creativity performance difference between machine and human designers. To achieve the goals, the study compared the E-creativity performance between humans and machine. A state-of-the-art model DALL • E is chosen to represent the E-creativity performance of computers while the creative outputs of novice designers are used to represent humans. The results reveal that the E-creativity level of machine is higher than that of humans. This result provides useful insights for supporting the development of the next-generation E-creativity engines for computational creativity.

2. THEORETICAL BACKGROUND

The creativity process is an idea-solution finding process [8, 9]. One way to understand creativity is to understand this searching process. In open-ended domains such as creativity, the iterative searching process can be termed as search strategy [10]. In this process, people move from one subset to another. The searching strategy can affect the creativity level of final solutions, although it is not observable directly. The outputs of each searching strategy stage are relied on to observe the searching strategy [9]. This searching strategy can also be applied in computational creativity to generate ideas, which is called exploratory creativity (E-creativity) in computational areas [11] as promoted by Boden [12] and then formalized by Geraint [13] and other researchers [6, 14, 15]. E-creativity is used to represent the process where people search ideas or artifacts within a given search space and are governed by certain rules [11, 12, 16]. In other words, E-creativity can represent how much creativity that the person applied in an iterative searching process. For example, if a person wants to design chips with a specific flavor. The process where the person based on the cucumber-flavor chips to generate the orange-flavor chips can be considered as an E-creativity process. The differences between E-creativity and creativity are also obvious. E-creativity is a form of creativity and is specifically used to represent the creativity of a person's searching strategy to generate more creative concepts.

Some tools, such as Ludoscope [5], "Black box" [9], "Narrative Search" [16], and "DeLeNoX" [11], have been developed based on E-creativity principles and applied in textual narrative or creative design. Dormans and Stefan [5] developed an algorithm to explain E-creativity in procedural content generation. In their study, the E-creativity process was transformed from combinational creativity and divided into two steps. They attempted to maintain the scalable and tractability of the combinational creativity and thus enable people to explore and establish the conceptual space. In this way, E-creativity can be achieved. In addition, to maintain the E-creativity suitable for the definition of creativity, their study also considered two attributes of creativity - novelty and usefulness, which results in the combinatorial logic algorithm and the reorganizing logic algorithm. This algorithm was applied in a maze game. The evaluation of the E-creativity performance is based on the relations between lock and keys. The evaluation aims to identify whether the algorithm can work more effectively than the previous version.

Another tool, "Black box", which is based on E-creativity, is developed by Kyle et al. [9]. They developed a computerized aesthetic composition task to capture people's E-creativity process. Then a theoretical model has been proposed. The model includes three phases: exploration phases, criteria description phases, and landscape rating phase. The researchers selected six aesthetic composition-related criteria to assess the E-creativity performance of this tool.

DALL•E is a program that aims to generate various images from textual descriptions. Aditya et al. [17] trained it based on the 12 billion parameter version of GPT-3. It can generate images according to the textual description that expresses various concepts. There are various approaches in DALL•E to generate images from text, including creating anthropomorphic versions of animals and objects, combining unrelated concepts reasonably, rendering text, and transforming existing images. Moreover, modifying a single attribute of the textual description can generate completely different images. For these text-to-image generation models, the model performance was assessed by identifying cosine distance between predicted style vectors on same or different style image pairs [18, 19].

DALL•E has an excellent capability in creating imaginary objects. The samples show that DALL•E can have a certain level of E-creativity through its "Drawing Multiple Objects" and "Inferring Contextual Details" capabilities. "Drawing Multiple Objects" allows DALL•E to control objects' attributes and spatial relationships simultaneously based on the given description and generate various images which are consistent with the description. "Inferring Contextual Details" allows DALL•E to draw the same object in different situations or generate an image of an object which includes a specific text. The two capabilities offer DALL•E the ability to generate different images when only part of the description is changed, which further gives DALL•E the chance to have a certain level of E-creativity. This also gives us the chance to compare the computational E-creativity and human E-creativity based on images. Therefore, the DALL•E was selected as the tool which can represent the computational E-creativity.

Among the review of existing computational E-creativity tools, existing research indicated some assessment criteria that may have the potential to be used to assess E-creativity. These criteria were often summarized from the definition of E-creativity [20] or what is to be E-creativity [21]. For example, Edward and Clifford [15] suggested that E-creativity is about the distance among the rules, characteristics, and structures used to generate concepts. Therefore, the distance in return can be used as the way to assess E-creativity. Rayasam [22] suggests that E-creativity represents whether the concepts have the same concept space or concept boundaries and thus the concept space may be the criteria to assess E-creativity. Till now, there is not a systematic and practical method that can be used to assess E-creativity.

To summarize, existing studies on computational E-creativity tools mainly focus on the mechanism of the tools and their effectiveness. However, whether the E-creativity performance of machine and human are the same have not been fully studied. This study, thus, takes the design process as an example to investigate the E-creativity performance differences between machine and human designers.

3. METHOD

As mentioned before, E-creativity is a kind of creativity. Considering that creativity can consist of person creativity, product creativity, press creativity, and process creativity, the E-creativity may also consist of person E-creativity, product E-creativity, press E-creativity, and process E-creativity. Since the product E-creativity is tangible in the form of images or text, product E-creativity was used to represent the E-creativity in this study. The whole study protocol was displayed in a form of a flow chart in Figure 1.

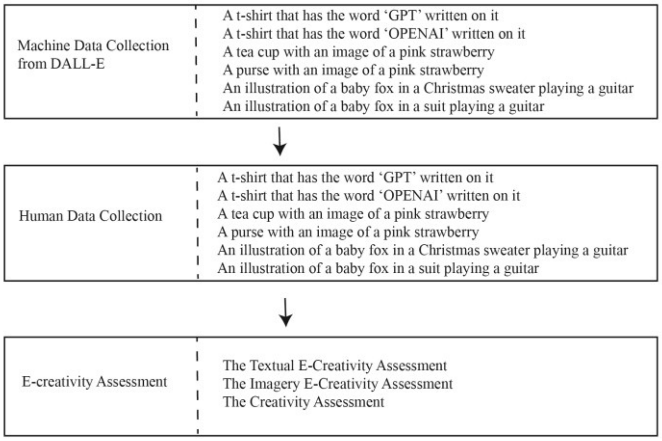

Machine Data Collection from DALL-E
A t-shirt that has the word 'GPT' written on it
A t-shirt that has the word 'OPENAI' written on it
A tea cup with an image of a pink strawberry
A purse with an image of a pink strawberry
An illustration of a baby fox in a Christmas sweater playing a guitar
An illustration of a baby fox in a suit playing a guitar
Human Data Collection
A t-shirt that has the word 'GPT' written on it
A t-shirt that has the word 'OPENAI' written on it
A tea cup with an image of a pink strawberry
A purse with an image of a pink strawberry
An illustration of a baby fox in a Christmas sweater playing a guitar
An illustration of a baby fox in a suit playing a guitar
E-creativity Assessment
The Textual E-Creativity Assessment
The Imagery E-Creativity Assessment
The Creativity Assessment

FIGURE 1: THE WHOLE STUDY PROTOCOL

3.1 Machine Data Collection

In this study, DALL•E is chosen as a representative for generating E-creativity imagery and is compared to novice designers. To express E-creativity more clearly, the study selected three groups of different textual descriptions patterns for the experiment. Each group has two sentences that can be used to describe E-creativity, which is differentiated by replacing a certain attribute or subject (Table 1). These textual descriptions are the input, while the outputs are imagery generated by DALL•E.

DALL•E initially create 512 plausible images from a sentence that explores the compositional structure of language. These images are ranked by the Contrastive Language-Image Pre-training (CLIP) method. Then, the top 30 images are displayed on the DALL•E official webpage. All of the sample images can be obtained from its official webpage: https://openai.com/blog/DALL•E/. The study selects the top five images for each task and uses them as test targets to represent the computational outputs of DALL•E.

In the process of selecting the images, three designers with more than four years of industrial design experience are recruited. Among the thirty images generated in each task, designers are asked to select the top five images from the sample pool as the machine dataset. The selection procedure is described as follows:

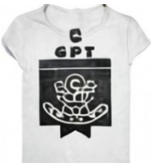
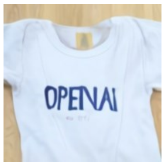
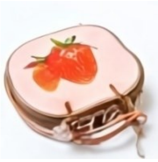
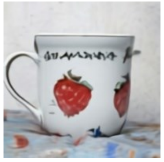
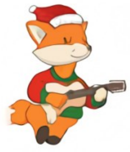
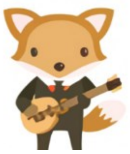
1) The images should qualify the corresponding textual description without obvious defects;

2) The images should have a clear background, such as white background;

3) The selection criteria are variety and fidelity. Variety means the selected images should be unique, novel, and non-repetitive as much as possible; Fidelity means the level of plausible and realistic should be as high as possible.

Two metrics are applied to determine the final five machine-dataset images for each task: i) voted by more than one designer; ii) higher ranking for a particular image. After the selection, all the selected images have been processed to the same resolution (256*256 pixels). The input textual descriptions and corresponding samples of the machine data are shown in Table 1 to give an overview of the machine data source.

TABLE 1: AN OVERVIEW OF MACHINE DATA

Group No.	Task No.	Textual Description	Design example
1	1	A t-shirt that has the word 'GPT' written on it	
	2	A t-shirt that has the word 'OPENAI' written on it	
2	3	A purse with an image of a pink strawberry	
	4	A tea cup with an image of a pink strawberry	
3	5	An illustration of a baby fox in a Christmas sweater playing a guitar	
	6	An illustration of a baby fox in a suit playing a guitar	

Since the 30 images are the top 30 images ranked by computer and then the experiment selects the top 5 images ranked by experienced designers, the selected five images are the images that both computer and human thought may have the potential to be the best ones. Therefore, the five images chosen from machine data may have a huge potential to represent the DALL•E's best E-creativity capabilities.

3.2 Human Data Collection

The subjects in comparison with DALL•E are novice designers. We recruited eight novice designers with less than four years of design experience to create a human dataset. The reason why novice designers instead of experts were used to represent the human designers is that novice designers tend to have less E-creativity. Since the aim of this study is to compare the E-creativity performance between machine and human. The start target group should be the machine and novice designers.

Eight designers gathered in a room with their laptops which had already been installed with computer-aided software and are ready to use. Before the experiment, an introduction to the task is given to the designers. Specifically, every designer is required to finish six graphic or product design tasks. In every task, designers need to complete a design based on textual descriptions which are the same as the machine datasets. In addition, designers are allowed to use any computer software which they are familiar with to produce a graph, such as Photoshop, and Adobe Illustrator.

TABLE 2: AN OVERVIEW OF HUMAN DATA

Group No.	Task No.	Textual Description	Design Example
1	1	A t-shirt that has the word 'GPT' written on it	
	2	A t-shirt that has the word 'OPENAI' written on it	
2	3	A purse with an image of a pink strawberry	
	4	A tea cup with an image of a pink strawberry	
3	5	An illustration of a baby fox in a Christmas sweater playing a guitar	
	6	An illustration of a baby fox in a suit playing a guitar	

Before the experiment, a pilot test was conducted to determine the specific requirements and the time that every task needs. According to the difficulty of the tasks, the first and second tasks are required to be completed within 20 minutes, while the third and fourth tasks are given 35 minutes and the last two tasks are given 50 minutes to complete. Participants can have a five-minute break between two design tasks to avoid fatigue on long-time design tasks. The experiment lasts four hours in total.

During each task, designers are allowed to use a single word to search on the internet for inspiration but are prohibited from searching with combined words. For example, the designers can search for 'cup' and 'strawberry' separately, but they are not allowed to search for patterns of cups with strawberries. Furthermore, the searched patterns can be used in the designers' design.

Others requirements regarding the design involve using a white background, which is in line with the image background of the machine dataset and thus the images will not be distinguished as human generated or computer created from the image background. The textual description that is not related to task requirements should not be included in the drawing. In addition, although participants are not told they need to display their E-creativity ability in these tasks, they have been announced that their design should maintain high quality as much as possible. The result of designs should be saved in 'jpg' or 'png' format, and the image quality of the design should be larger than 512*512 pixels.

To ensure the quality of the human dataset, three professional designers with master's degrees in industrial design were employed to select the top 5 images with the highest quality. The selection procedure is the same as the machine dataset. Finally, thirty images (6 sets * 5 images) were obtained as a human dataset for the following experiment, and then these images were converted to the same resolution (256 * 256 in pixel). The textual descriptions and corresponding examples of the six sets of designs are shown in Table 2.

3.3 E-creativity Assessment

This research concerns whether DALL•E can achieve E-creativity at the human level. Hence, we utilize expert tests to deeply investigate this and provide interpretable results. The assessment included three parts: textual E-creativity assessment, imagery E-creativity assessment, and creativity assessment. The E-creativity assessment was performed by showing the definition of E-creativity to assessors, and assessors need to follow the definition to self-determine the E-creativity levels of the given text or images. The creativity is assessed based on the novelty, feasibility, and completeness of the given graphs.

The expert test was designed and conducted via an online questionnaire built up by Qualtrics. The usability and feasibility of this website are thoroughly tested to ensure that the questionnaire can be displayed correctly on both PC and mobile phones.

At the beginning of the online questionnaire, the instructions and requirements related to the test are presented on the homepage, and then experts are required to fill in their age, gender, profession, educational background, and the time of design training or education so that demographic information can be counted. The expert test includes three parts: textual E-creativity assessment, creativity assessment, and images E-creativity assessment. Before the formal evaluation, a pilot test is conducted to refine the specific requirements and check whether there are other problems within the questionnaire.

Twelve designers, who are all students or practitioners with at least three years of design education, are employed for our expert test. The collected numbers of the two different questionnaires are the same. Each questionnaire is finished by six participants. The reason why three professional designers were used to down-select DALL•E images, whereas 12 expert designers with 3 years of design education were used to rate creativity and E-creativity is that the down selection task is a heavy workload task. The results were also hard to get consistency. Therefore, few professional people were recruited in this process to ensure high consistency. As for the creativity and E-creativity assessment task, it is a low workload task. The results were also easier to get a consistency. Therefore, more participants were recruited to ensure the reliability of this assessment. The number of assessors was determined from existing research of Tarricone and Newhouse [23].

The specific evaluation of each part is explained in turn as below:

The Textual E-Creativity Assessment. Both DALL•E and the human designers generated their design based on the text prompt, thus it is worth detecting relations between textual and imagery E-creativity and further identifying whether these relations are different between humans and machines. In addition, the study focuses on product E-creativity. Text and images are the two forms of product E-creativity. Therefore, the textual E-creativity is also worth assessing. In other words, evaluating textual E-creativity is to give a quantitative reference of the selected textual description when compared to imagery E-creativity.

In this assessment, the definition of E-creativity was first given. Since the general definition from existing research may be difficult to be understood by participants. An easier version was given. To be specific, "E-creativity is a new idea B inferred from a known idea A according to a certain paradigm (or law). This kind of "thinking" or "reasoning" can be considered E-creativity".

Also, an example of textual E-creativity was given. To be specific, if the participants want to assess the textual E-creativity between "a cartoon penguin with red hat" and "a cartoon penguin with blue hat", the assessors need to first identify how "a cartoon penguin with blue hat" inferred from "a cartoon penguin with red hat" according to a certain paradigm (or law). Then, this kind of "thinking" or "reasoning" can be considered textual E-creativity of this task.

At the beginning of the test, this definition of E-creativity and corresponding samples are given to participants, who are required to maintain a consistent understanding of the definition throughout the assessment. Participants are asked to assess the textual E-creativity performance of two textual descriptions which are in the same group. The score from 1 to 5 indicates a gradual increase in E-creativity level.

The Creativity Assessment. Consensual Assessment Technique (CAT) is introduced as the creativity assessment method. CAT is a method developed by Amabile [24] and is usually used for creativity assessment. It has two steps: ask participants to create some design ideas, and then ask experts in the domain to evaluate the creations using a Likert-type scale. In our creativity assessment, three indicators of novelty, feasibility, and completeness are used to measure the quality of a single image produced by human designers or machine. Novelty refers to whether the design is uncommon, original, and attractive; Feasibility refers to whether the idea is in line with common sense, natural, and coordinated; Completeness refers to whether the creativity meets the textual description and whether the quality of the image is at a high level.

Each page has only one image given to the participants on the top of the page with a corresponding textual description to maintain the cleanliness and readability of the page. In addition, whether the images were from humans or DALL•E is not revealed to experts in the evaluation. Participants are asked to measure the image's novelty, feasibility, and completeness using a Likert-type scale. All the images are presented in a random order.

The Imagery E-Creativity Assessment. Imagery E-creativity is evaluated to investigate the performance gap between humans and DALL•E. The E-creativity of the three groups in both datasets is assessed separately. The evaluation of each group is performed in an independent webpage. In each group section, each set consists of two images that are respectively selected from the two textual descriptions within the same group. Therefore, a total of ten sets of data are sorted within each page. To be specific, five sets of images are selected from human outputs and five sets of images are selected from machine outputs. The order of the ten sets is random. In each group, there are no paired relations for the DALL•E generated images while a pair of images could be produced by the same designer. To avoid the impact of the no fixed paired relations in the machine dataset, two questionnaires with different machine image paired relations are made and randomly distributed. Assessors are not informed whether each set of ideas is generated by humans or DALL•E.

In this assessment, the definition of E-creativity was first given again to remind assessors what E-creativity is. Also, an example of imagery E-creativity assessment was given. To be specific, if the participants want to assess the imagery E-creativity between the image generated from "a cartoon penguin with red hat" and the image generated from "a cartoon penguin with blue hat", the assessors need to first identify how the image generated form "a cartoon penguin with blue hat" inferred from the image generated form "a cartoon penguin with red hat" according to a certain paradigm (or law). Then, this kind of “thinking” or “reasoning” can be considered imagery E-creativity of this task.

The textual description of each group is displayed at the beginning of the page, for example, "the idea on the left is based on a T-shirt that has the word 'GPT' written on it, and the idea on the right is based on a T-shirt that has the word 'OPENAI' written on it." Assessors are asked to rate the E-creativity levels between the two images in each set based on their understanding of what E-creativity is.

4. RESULTS

4.1 Results of Textual E-Creativity Assessment

To identify the textual E-creativity, mean values and the standard deviation (SD) of the each-group textual E-creativity are calculated. The results are reported in Table 3. From the results, it could be found that the average textual E-creativity score of Group 2 is 2.11 (SD = 0.99) which is the same as that of Group 3 (SD = 0.88). The average textual E-creativity score of Group 1 is 2.16 (SD = 1.30) which is higher than Group 2 and Group 3. However, the difference of the textual E-creativity score is only 0.05 (about 1%), which is not obvious. The inter-rater reliability of three groups was all in the moderate levels (0.43, 0.41, and 0.46 respectively)

TABLE 3: THE TEXTUAL E-CREATIVITY RESULTS OF THREE GROUPS

Group No.	Textual E-creativity score	SD	Inter-rater Reliability
1	2.16	1.30	0.43
2	2.11	0.99	0.41
3	2.11	0.88	0.46

4.2 Results of Imagery E-Creativity Assessment

The imagery E-creativity results of humans and machine are shown in Table 4. The results show that in Group 1, the performance of imagery E-creativity on humans is better than that of the machine while in Group 2 and Group 3, the performance of imagery E-creativity on the machine is better than that of humans. To verify whether the results are statistically significant, a paired samples test was conducted. The results reveal that when considering the three groups as a whole group, the performance of imagery E-creativity on the machine (2.47) is better than that of humans (2.03). This result is statistically significant (p=.022<.05). The inter-rater reliability is mainly located in the moderate agreement or substantial agreement levels.

TABLE 4: THE IMAGERY E-CREATIVITY RESULTS OF HUMAN AND MACHINE

Group No.	Dataset	1	2	3	4	5	Mean	SD	Inter-rater Reliability
1	Human	3.00	2.85	3.00	3.08	3.15	3.02	0.10	0.46
	Machine	3.77	1.77	3.00	2.31	2.15	2.60	0.71	0.69
2	Human	1.85	1.62	2.08	1.31	2.62	1.89	0.44	0.53
	Machine	3.08	2.92	2.31	2.46	2.62	2.68	0.29	0.68
3	Human	1.85	1.69	2.54	2.23	2.00	2.06	0.30	0.63
	Machine	3.54	2.85	3.15	2.92	3.38	3.17	0.26	0.65
	Human	/	/	/	/	/	2.03	0.91	
	Machine	/	/	/	/	/	2.47	0.98	

4.3 Results of Creativity Assessment

The creativity was assessed through the following three metrics: novelty, feasibility, and completeness. The results are reported in Table 5. The results reveal that the creativity level of novice designers is higher than that of machine regarding novelty, feasibility, and completeness. To verify whether the results are statistically significant, the paired samples test was conducted. The results reveal that when considering the three groups as a whole, the novelty, feasibility, and completeness of humans (2.52, 3.11, and 3.05 respectively) are higher than that of machine (2.07, 2.72, and 2.41 respectively). These differences are statistically significant (p=.000<.05). The inter-rater reliability is mainly located in the substantial agreement level.

TABLE 5: CREATIVITY ASSESSMENT RESULTS ON HUMAN AND MACHINE

Criterion	Dataset	Mean	SD	Inter-rater Reliability
Novelty	Human	2.52	1.04	0.72
	Machine	2.07	0.83	0.63
Feasibility	Human	3.11	1.20	0.71
	Machine	2.72	1.08	0.64
Completeness	Human	3.05	1.19	0.53
	Machine	2.41	0.98	0.61

5. DISCUSSION

5.1 The Performance between DALL•E and Designers Regarding E-Creativity

The textual E-creativity is similar among the three groups. This may reflect that the E-creativity in the three groups' texts is similar. The E-creativity score of each pair is all above two. Since E-creativity is reflected by the exploratory reasoning capability, the low score indicates that the three pairs of text are not perceptually difficult to evolve from one to the other. It also reveals that the text pattern is easier to realize than in the imagery form. However, the E-creativity is evaluated by the perceptual exploratory reasoning capability and few findings are supporting the effectiveness of this method. Therefore, further research on the evaluation of E-creativity is expected.

The imagery E-creativity performance of DALL•E is higher than that of novice designers. Limited by their visual expression capabilities, the novice designers may not be able to have a better imagery reasoning capability as semantic reasoning when compared to DALL•E. There is another finding from the human dataset, in which designers tended to work in an easy and less workload way to complete the task. As a result, similar graphic elements appeared in the same group. As for the machine (DALL•E), its generation capability is completely based on the arithmetic which is designed to achieve a certain level of variety. Therefore, DALL•E does not have the trend to generate similar elements in the same group. This difference suggests that DALL•E has a better performance on imagery E-creativity than humans. This result indicates that DALL•E is better at transforming textual E-creativity to imagery creativity than novice designers, and it can assist designers to generate more creative ideas.

5.2 The Comparison between Textual and Imagery E-Creativity Performance

When comparing the imagery E-creativity with textual E-creativity, it could be found that the performance of imagery E-creativity is higher than that of textual E-creativity in terms of DALL•E; while the novice designers' performance of imagery E-creativity is lower than that of textual E-creativity.

One explanation could be that in the process of the textual E-creativity assessment, assessors did not need to consider the corresponding visualized outputs. Therefore, their justification is completely based on semantic reasoning knowledge. For each group's textual descriptions, as shown in Table 1 and Table 2, only one or two attributes change. Based on the semantic reasoning knowledge, the textual E-creativity was evaluated in low or middle levels. As for the imagery E-creativity assessment, it is based on visual reasoning. As explained before, designers tend to produce similar graphic elements in the same group, showing a certain level of design fixation, while DALL•E appeared to have a certain level of design innovation [25].

When considering each group's result separately, it could be found that there is little difference among the three groups regarding textual E-creativity, while the variation is large for imagery E-creativity no matter which dataset it is. It might reveal that the perception of semantic E-creativity is less various due to its abstraction from the cognitive perspective, compared to visual E-creativity [14]. In particular, the visual imagination of textual E-creativity could produce a much large space to explore which might result in the imagery E-creativity evaluation scores with high variety.

5.3 The Comparison between Creativity and Imagery E-Creativity

The creativity score of humans is higher than that of machine; while the imagery E-creativity performance of humans is not as good as that of machine. This result indicates that E-creativity and creativity may have less correlation. It can be explained by the different working mechanism of both concepts. In the E-creativity process, humans or machine tend to transcend the limitation of a set of fixed rules [26]. In the creative process, human tend to come up with novel and valuable concepts [27]; while machine tends to generate concepts that observers think is creative [28]. The different thinking process leads to the less correlation between E-creativity and creativity are less correlated. Another explanation can be given from the nature of E-creativity and creativity concepts as E-creativity reflects the exploratory reasoning capability while creativity centers on novelty, feasibility, and completeness. The different focuses lead to the difference between E-creativity and creativity.

5.4 Limitation and Future Research

In this study, limited by the existing research on E-creativity evaluation, E-creativity is reflected by the perceptual exploratory reasoning capability. The higher E-creativity may lead to a higher exploratory reasoning capability. However, a higher exploratory reasoning capability may not certainly lead to a higher E-creativity. In other words, there might be other better metrics measuring the E-creativity performance and the exploratory reasoning capability can be only one metric of them. Therefore, more thorough research on the measurement of E-creativity is necessary. In addition, DALL•E is a machine learning model which is mainly used for transforming texts to images. In other words, its core arithmetic is not for simulating E-creativity. This means that DALL•E may not be the best model used to produce E-creativity exclusively. Therefore, a computational E-creativity generation engine is expected in future research.

6. CONCLUSION

The study compared the E-creativity performance between a computational model and human designers. The study involved two datasets. The machine dataset is made by DALL• E, a model that can transform the text into an image. The human dataset is generated by designers based on the same textual cues used in the machine dataset. Experts as raters are invited to assess the textual description E-creativity, imagery E-creativity, and creativity between the machine and humans. The results reveal that the creativity performance of humans is higher than that of the machine, while the E-creativity of the machine is higher than that of humans. The textual E-creativity is higher than the imagery E-creativity performance of humans while it is lower than imagery E-creativity performance of the machine. The analysis further provides insights for supporting the development of more advanced E-creativity engines and corresponding evaluation methods.

ACKNOWLEDGEMENTS

This paper is funded by National Key R&D Program of China (2018AAA0100703), and the National Natural Science Foundation of China (No. 62006208 and No. 62107035).

REFERENCES

[1] Han, Ji, Forbes, Hannah and Schaefer, Dirk. “An exploration of how creativity, functionality, and aesthetics are related in design.” *Research in Engineering Design* Vol.32 No.3 (2021):pp. 289-307.10.1007/s00163-021-00366-9

[2] Israel-Fishelson, Rotem, Hershkovitz, Arnon, Eguíluz, Andoni, Garaizar, Pablo and Guenaga, Mariluz. “The associations between computational thinking and creativity: The role of personal characteristics.” *Journal of Educational Computing Research* Vol.*58* No.8(2021):pp.1415-1447.10.1177/0735633120940954

[3] Mateja, Deborah and Armin Heinzl. “Towards Machine Learning as an Enabler of Computational Creativity.” *IEEE Transactions on Artificial Intelligence* Vol.2No.6 (2021):pp. 460-475. 10.1109/TAI.2021.3100456

[4] Moruzzi, Caterina. “Measuring creativity: an account of natural and artificial creativity.” *European Journal for Philosophy of Science* Vol.11 No.1 (2021):pp. 1-20.10.1007/s13194-020-00313-w

[5] Dormans, Joris and Stefan, Leijnen. “Combinatorial and exploratory creativity in procedural content generation.” *fourth workshop on Procedural Content Generation for Games at the Foundations of Digital Games Conference*:pp. 1-4.Chania, Greece, May 14-17, 2013.https://hdl.handle.net/2066/123037

[6] Cardoso F, Amílcar. “Autonomous Composition as Search in a Conceptual Space: A Computational Creativity View.” *International Symposium on Computer Music Multidisciplinary Research*:pp. 187-198. Springer, Cham, 2017. 10.1007/978-3-030-01692-0_13

[7] Bergs, Alexander and Nikola, Anna Kompa. "Creativity within and outside the linguistic system." *Cognitive Semiotics* Vol. 13 No.1 (2020): pp. 2020-2025.10.1515/cogsem-2020-2025

[8] Howard, Thomas, Stephen J, Culley and Elies, Dekoninck. “Creativity in the engineering design process.” *DS 42: Proceedings of ICED 2007, the 16th International Conference on Engineering Design*:pp. 329-330.Paris, France, August 28 - 31, 2007.

[9] Kyle E, Jennings, Dean Keith, Simonton and Stephen E, Palmer. “Understanding exploratory creativity in a visual domain.” *Proceedings of the 8th ACM conference on Creativity and cognition*:pp. 223-232. November, 2011.10.1145/2069618.2069656

[10] Gubenko, Alla, Kirsch, Christiane, Smilek, Jan Nicola, Lubart, Todd and Houssemand, Claude. “Educational Robotics and Robot Creativity: An Interdisciplinary Dialogue.” *Frontiers in Robotics and AI*, Vol.8, No.178. (2021).10.3389/frobt.2021.662030

[11] Antonios, Liapis, Hector P, Martinez, Julian Togelius and Georgios N, Yannakakis. “Transforming exploratory creativity with DeLeNoX.” *Proceedings of the Fourth International Conference on Computational Creativity*:pp.56-63. 2013.arXiv preprint arXiv:2103.11715

[12] Boden, Margaret. *The creative mind: myths and mechanisms Weidenfeld*. Abacus & Basic Books 4 (1990).

[13] Geraint A, Wiggins. “Searching for computational creativity.” *New Generation Computing* Vol.24 No.3(2006):pp.209-222. 10.1007/BF03037332

[14] Duch, Wlodzislaw. “Computational creativity.” *The 2006 IEEE International Joint Conference on Neural Network Proceedings*: pp. 435-442.Vancouver, BC, Canada, July 16-21, 2006.10.1109/IJCNN.2006.246714

[15] Edward CK, Hung, and Clifford ST, Choy. “Conceptual Recombination: A method for producing exploratory and transformational creativity in creative works.” *Knowledge-Based Systems* Vol. 53 (2013):pp. 1-12.10.1016/j.knosys.2013.07.007

[16] Mark O, Riedland Michael R., Young. “Story planning as exploratory creativity: Techniques for expanding the narrative search space.” *New Generation Computing* Vol.24 No.3(2006):pp.303-323.10.1007/BF03037337

[17] Aditya, Ramesh, Mikhail, Pavlov, Gabriel, Goh, Scott, Gray, Chelsea, Voss, Alec, Radford, Mark, Chen and Ilya Sutskever. “Zero-shot text-to-image generation.” *International Conference on Machine Learning*:pp. 8821-8831. PMLR, July, 2021.

[18] Yufan, Zhou, Ruiyi, Zhang, Changyou, Chen, Chunyuan, Li, Chris, Tensmeyer, Tong, Yu, Jiuxiang Gu, Jinhui, Xu and Tong, Sun “LAFITE: Towards Language-Free Training for Text-to-Image Generation.” (2021). arXiv preprint arXiv:2111.13792

[19] Christoph, Schuhmann, Richard Vencu, Romain, Beaumont, Robert , Kaczmarczyk, Clayton Mullis, Aarush, Katta, Theo, Coombes, Jenia Jitsev and Aran, Komatsuzaki. “Lion-400m: Open dataset of clip-filtered 400 million image-text pairs.” (2021). arXiv preprint arXiv:2111.02114

[20] Weiss, Selina and Oliver, Wilhelm. “Coda: Creativity in psychological research versus in linguistics–Same but different?.” *Cognitive Semiotics* Vol.13 No.1(2020).10.1515/cogsem-2020-2029

[21] Anna, Jordanous. “A standardized procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative.” *Cognitive Computation* Vol.4 No.3(2012):pp. 246-279.10.1007/s12559-012-9156-1

[22] Rayasam, Sushma. *Transformational creativity in requirements goal models* . Master's thesis. University of Cincinnati. Cincinnati, US. 2016. http://rave.ohiolink.edu/etdc/view?acc_num=ucin1471346090

[23] Tarricone, Pina., and Newhouse, C. Paul. An investigation of the reliability of using comparative judgment to score creative products. *Educational Assessment* Vol.22 No.4(2017):pp. 220-230.https://doi.org/10.1080/10627197.2017.1381553

[24] Teresa M, Amabile. "Social psychology of creativity: A consensual assessment technique." *Journal of personality and social psychology* Vol.43 No.5 (1982):pp. 997–1013.10.1037/0022-3514.43.5.997

[25] Thanh An, Nguyen and Yong Zeng. "A theoretical model of design fixation." *International Journal of Design Creativity and Innovation,* Vol.5 No.3-4 (2017):pp. 185-204.10.1080/21650349.2016.1207566

[26] Ławrynowicz, Agnieszka. "Creative AI: A new avenue for the Semantic Web?." *Semantic Web* Vol.11 No.1 (2020):pp. 69-78. 10.3233/SW-190377

[27] Boden, Margaret. "Understanding creativity." *Revolutionary Changes in Understanding Man and Society*:pp. 75-82. Springer, Dordrecht, 1995.10.1007/978-94-011-0369-5_5

[28] Colton, Simon and Geraint A, Wiggins. "Computational creativity: The final frontier?."*ECAI 2012: 20th European Conference on Artificial Intelligence*:pp. 21-26. Montpellier, France, August 27-31, 2012.10.3233/978-1-61499-098-7-21