Human Judgement of AI-generated Art Content

Ewelina Gajewska

ewegaj@st.amu.edu.pl

Abstract

In his 1950 publication Alan Turing proposed a test for machine intelligence known later under the name of the Turing Test (TT). It states that if a human judge could do no better than random guessing at determining a machine in a human-machine conversation, then the machine passes TT and thus, it could be called intelligent. Since its introduction in 1950s several versions of TT have been proposed, including the one inspired by the Lady Lovelace objection against the original TT. Lady Lovelace claimed that a machine cannot be regarded as intelligent unless it has the capacity to produce something novel, creative like a piece of art. This modified version of TT called the Lovelace Test of intelligence, proposed in (Bringsjord et al., 2001) and later updated in (Riedl, 2014), is employed in the current study to test two hypotheses regarding machine intelligence. The first states, following the Turing's idea, that a human judge could do no better than chance at determining a machine-generated artwork from the one created by a man. The second one regards aesthetic value of the AI- and human-generated images – that the former category of images are rated as aesthetically pleasing as the latter. Hypotheses are tested in an empirical study – online questionnaire.

Keywords: artificial intelligence, computational creativity, Lovelace Test, Turing Test

1 Introduction

In 1950 Alan Turing posed a question "Can machines think?" (Turing, 1950) which started a discussion about artificial intelligence still relevant today. While the question itself was too meaningless to deserve considering for Turing, the researcher proposed the idea of the imitation game, known later under the name of the Turing Test (TT). Shortly after the Turing's publication, TT started to be viewed as the ultimate test of machine intelligence and is regarded as such nowadays (French, 2012).

Rules of the imitation game are as follows. We have 2 participants – a man and a woman – and an interrogator (judge) who asks a series of questions which both individuals need to answer. The object of the game is to tell who is a man and who is a woman based on the replies provided by both participants. A judge and participants, in addition, stay in separate rooms and the answers are written down on a paper, so that only the content of responses from participants are evaluated.

Then, one of the participants is replaced with a machine and the procedure of interrogation is repeated. At the end of each round of questions, a judge evaluates both responses and decides which of the two participants is a machine and which of them is a man. So, the role of the machine is to imitate a human behaviour to the greatest extent, including making mistakes such as spelling errors and adjusting the speed of answering a question (if the answers are supplied in real time).

So the question here is whether a judge will achieve the same (low) accuracy of recognition of a man and a machine as in the case of identification of a man and a woman. If so, one could conclude a machine is able to deceive a human on a certain level and thus, is capable of imitating "human machines", including a part of their intelligent behaviour. What is more, Turing gave a prediction that by the turn of the century in a five minute human-machine interaction (in a linguistic form; via email, for example) a lay person would have no more than a 70% chance of making a human/machine determination (Turing, 1950, p. 442).

In fact, Turing proposed two versions of his imitation game. In his 1950 publication (Turing, 1950) the researcher describes the idea of imitation game with three participants (two tested subjects and a judge), as well as the other version of the game, in which only two participants are involved (one contestant and one judge). The former is referred to as *parallel-paired* Turing Test, while the latter is recognised as the *viva voce* Turing Test. Both TT setups are utilised in the current study.

Bringsjord et al. (2001) propose an alternative Turing Test of intelligence – the Lovelace Test. The name comes from the Lady Lovelace who believed that a machine could produce only what it ordered to do. Following this idea, one might say that a machine could be called intelligent (in a human-like manner) only if it has the ability to originate things (a work of art or other creative concept). The argument formulated by Lady Lovelace was one of the objections articulated against the Turing's idea of the imitation game.

The authors argue that original TT falls short as a test of intelligence because of issues raised by the well-known Chinese Room Argument (Searle, 1980). That is, a machine is designed to mindlessly follow a strict rulebook. Bringsjord et al. (2001) in turn argue for a certain epistemic relation between a an artificial agent a, its output o, and the human architect d. In their proposition of a modified TT (the Lovelace Test) the authors say that artificial agent a, designed by d, passes the test if:

- a outputs o;
- o is the result of processes a could repeat;
- d cannot explain how a produced o.

Riedl (2014) proposes the updated Lovelace Test which is designed to be employed as a test of human-level intelligence for artificial intelligence. Although the test seems to be formulated in a way so as to test the creative ability of AI systems, Riedl (2014) argues that this ability requires a wide range of human-level capabilities and therefore could be regarded as a test of human-level intelligence for AI systems. In order to pass this version of the test, four conditions must be fulfilled:

- artificial agent a needs to create an output o of type t;

- output o complies with a set of criteria C that are expressible in natural language;
- a human judge j, having chosen t and C, is satisfied with o as a valid instance of t that meets C;
- a human referee r determines that the combination of t and C as being realistic for an average human.

At the same time, the author defines computational creativity as "the art, science, philosophy, and engineering of computational systems that, by taking on particular responsibilities, exhibit behaviours that unbiased observers would deem to be creative" (Riedl, 2014, p. 2). The essence of the updated Lovelace Test is the fulfilment of a set of constraints required by a human judge – if a machine is able to respond to those constraints no worse than an average person, it is compelling evidence of intelligence.

In turn Boden (2010) propose two alternative criteria for passing the Turing Test in the domain of machine-generated art. "I will take it that for an 'artistic' program to pass the TT would be for it to produce artwork which was: (1) indistinguishable from one produced by a human being; and/or (2) was seen as having as much aesthetic value as one produced by a human being" (Boden, 2010, p. 409). Further Boden (2010) writes that in order to classify art content as generated by a machine it must be independent of direct intervention by a human subject. However, Pease and Colton (2011) criticise the idea because of a lack of emphasis on the interactional aspect of the original TT and distinction between human and machine-produced artworks.

Novel AI systems such as DALL-E 2 developed by OpenAI[1] could account for the first objection – interactivity in the modified TT. DALL-E 2 can create realistic images and art from a so-called text prompt, i.e. description in natural language. A human judge could, therefore, formulate specific criteria for the image that AI (and a human artist) must produce. Then, if a machine and a human artist are provided with the same instructions (analogous to questions asked by a judge in the original TT) both artworks can be directly compared and evaluated by the judge.

Carnovalini and Rodà (2020) mention two other evaluation approaches based on the concept of Turing Test. First is the Consensual Assessment Technique proposed originally by Teresa Amabile as a method of evaluation of human creativity which, can be used also for evaluation of human vs. AI-generated artworks. In the second approach, a machine is provided with a material that needs to be learned from. Then, a set of instances generated by the machine are evaluated by a group of human judges along with the original content in the corpus. The task of a judge here is to decide whether a music composition, for example, was made by a human or a machine.

The present work asks several questions. Can people reliably distinguish AI-generated and man-made artworks? How accurate are their predictions? Are there differences in the aesthetic evaluation of AI-generated and man-made artworks? Finally, can machines pass a modified (Lovelace-like) Turing Test? The aim of the current work is to test two hypotheses regarding "machine intelligence". First (H-1), naive participants are not able to accurately tell AI- and human-generated art content (images) apart. Second (H-2), there are no significant differences in their judgements of aesthetic value between AI-produced and man-made paintings.

2 Related Work

The work of (Hong and Curran, 2019) examined human perception of artwork generated by artificial intelligence (AI). Specifically, the researchers tested how the knowledge about the author identity (machine vs. human) affects human evaluation of art. Judgements of artistic value were not equivalent between artwork produced by AI and humans – the latter condition gather higher scores on the artistic value scale. Composition, degree of expression, and aesthetic value are the three variables that with significantly higher scores of human-created artwork regardless of attributed identity of the artist.

[1] `https://openai.com/dall-e-2/`

However, evaluations on another variable – development of personal style – showed significant differences between conditions with AI and human attributed artist identity. In addition, Hong and Curran (2019) conclude that a negative attitude toward AI-created art strongly influences the evaluation of artworks when people believe they were created by AI.

The work by Ragot et al. (2020), in turn, show a negative bias in the public perception towards art made by AI. In their study, paintings presented as created by humans were evaluated significantly higher in terms of four subjective dimensions: liking, beauty, novelty, and meaning. A questionnaire with Likert scales was designed for this purpose. Regarding all four variables – declared liking, perceived beauty, novelty and meaning – results showed the main effect of induction (AI vs. human), the type of painting (landscape vs. portrait) and the real author (AI vs. human) with small effect sizes (measured by Cohen's d). In an additional recognition task, paintings made by humans were correctly recognised in 66% of cases and AI-generated artwork was correctly recognised in 56% of cases. Moreover, recognition rate of the authors of paintings was higher for portraits than for landscapes (69% and 53%, respectively).

Chamberlain et al. (2018) investigate the public response to visual art created by by humans and computers. Furthermore, the authors tested participants ability to discriminate between computer-generated and man-made art, and a potential bias towards the former category. The authors obtain similar results to Ragot et al. (2020) in terms of prejudice and negative bias towards AI-generated art. Analysis of results reveals it is driven mostly by the belief of machines' limited abilities regarding creativity. The authors provide also an interesting remark regarding Turnig-like tests of intelligence. "An important and understudied psychological question relating to this phenomenon is the extent to which individuals are willing to accept computer art as having the same worth and aesthetic value as that of a human artist, regardless of whether it passes such stringent tests of human-level intelligence" (Chamberlain et al., 2018, p. 178).

Köbis and Mossink (2021) assessed in a TT-like fashion whether people are able to distinguish algorithm-generated versus human-written poems, as well as preference of algorithm-generated versus human-written text. The authors employed GPT-2 for the purpose of their study – state-of-the-art Natural Language Generation algorithm (Radford et al., 2019). A text prompt comprised several lines of human-written poems that an algorithm needed to complete. There were several iterations of poem generations and either a random example (human-out-of-the-loop) or the best one (human-in-the-loop) was chosen for the study. Participants could not reliably distinguish human-written and machine-generated poems in the latter condition. Although, participants slightly preferred human poems regardless of whether they were aware of the algorithmic origin of the text or not.

In the related area of deepfakes (hyper-realistic manipulations of audio-visual content) detection, the study by Köbis et al. (2021) show that human participants could not reliably differentiate deepfake from authentic videos (57.6% accuracy rate on average), and at the same time they overestimate their own abilities to do so.

The current study follows previous works in the field, presented above. My study investigates (differences in) human judgement of aesthetic value regarding images generated by AI and humans, similarly to (Hong and Curran, 2019; Ragot et al., 2020). However, it goes beyond these works as they focus on the investigation of fine-grained subjective dimensions of aesthetic judgements and/or influence of a reported identity of an artist on those judgements (studying prejudice/bias towards AI). On the other hand, the present study examines humans' capacity to distinguish man-made and AI-generated images with a twofold methodology. First, it follows the protocol of a parallel-paired TT (modified into a Lovelace-like TT). Second, it implements a *viva voce* version of TT with a single subject being examined (interrogated) at a time. Thus, my study also allows to compare 'levels of difficulty' of different versions of TT.

3 Method

3.1 Material

Material for the study comprise man-made paintings available on artists' public profiles (personal websites, Instagram and Pinterest accounts), and AI-generated images posted by the owner of DALLE system – OpenAI on platforms such as Instagram and Twitter. DALLE is used as the only source of AI-generated images. There are, however, numerous human artists whose paintings are utilised in the study. The reason is that it is currently a state-of-the-art system for image generation similarly to Radford et al. (2019) who utilised only one language model – GPT-2 and poems from various human writers. Therefore, an image generated by DALLE was chosen first and then, a comparable painting in terms of style and content was selected from a human artist. This methodology allows to regard DALLE as a machine that takes part in TT.

Figure 1: Images chosen for the task of aesthetic value judgement in the study. Images A1 and A3 are generated by a machine, whereas paintings A2 and A4 are created by human artists. **A1**. 'Théâtre D'opéra Spatial' AI-generated painting by Jason M. Allen. **A2**. 'Interior of the Salon of the Archduchess Isabella of Austria' by Willem van Haecht. **A3**. 'Dino party' generated by the OpenAI system DALLE. **A4**. 'Surrounded by color' by Mike Winkelmann.

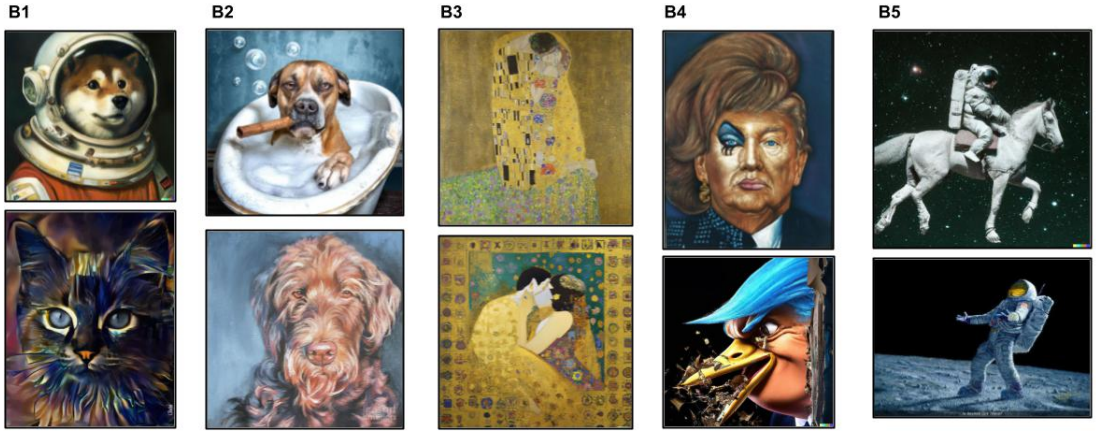

Figure 2: Images chosen for the parallel-paired modified TT task. Top images in pairs B1, B2 and B5, and bottom images in pairs B3 and B4 are AI-generated. **B1**. Top: 'Portrait of a shiba inu astronaut, oil painting, 16th century' AI-generated image by the OpenAI system DALLE. Bottom: 'Cat' by Léa Roche. **B2**. Top: 'A dog taking a bath with bubbles in an old bathtub and smoking a cuban cigar' generated by the OpenAI system DALLE. Bottom: 'Labradoodle' by Lee Ann Shepard. **B3**. Top: 'The Kiss' by Gustav Klimt. Bottom: Transformation of Gustav Klimt's 'The Kiss' generated by the OpenAI system DALLE. **B4**. Top: 'Donald Trump in Drag Dragrace' by Alfredo (Argo) Rodriguez. Bottom: 'Donald Trump' generated by the OpenAI system DALLE. **B5**. Top: 'A photo of an astronaut riding a horse' generated by the OpenAI system DALLE. Bottom: 'Is-Anyone-Out-There' by Alan Bean.

Figure 3: Images chosen for the *viva voce* modified TT task. Images C2 and C3 are generated by an AI system. **C1**. 'Slow and steady' by Mike Winkelmann. **C2**. 'Teddy bears mixing sparkling chemicals as mad scientists' by the OpenAI system DALLE. **C3**. 'Golden retriever puppy sitting at a diner drinking a cup of coffee, looking melancholy, edward hopper' generated by the OpenAI system DALLE.

3.2 Questionnaire

A modified Turing Test was used to investigate human capacity to distinguish man-made and AI-generated artworks. In addition, the study examined human judgements of aesthetic value of artworks (images) generated by a human artist or by a machine. The questionnaire comprises 12 questions in total. These questionnaire items could be grouped into 3 different categories of questions or 3 different tasks. The original questionnaire is available in Google Forms[2].

First, participants are asked to judge the aesthetic value of each of 4 paintings presented in Figure 1 – 2 generated by AI and 2 created by a human artist. Five-point Likert scale is used for this task with option '1' meaning 'I do not like this image' and option '5' meaning 'I like this image very much'[3].

The second type of questions is a modified parallel-paired Turing Test. The original form of the question-response pair in TT is replaced by the Lovelace-like test of intelligence. That is, participants of the study play a role of judges (interrogators) in TT. They are presented with a pair of images – one generated by a machine and the other created by a human. The task of a judge here is to choose either image 1 or image 2 as the one generated by AI[4]. There are 5 pairs of images each participant has to choose from, which are presented in Figure 2. In addition, participants were asked to shortly justify their choice in an open question form.

The third category of questions is designed in a *viva voce* TT fashion. That is, similarly as in the previous type of questions, human participants play a role of judges (interrogators) in the Lovelace-like TT. However, this time each participant is presented with a single image generated either by a human or by AI and asked whether in their opinion this image was generated by AI[5]. There are 3 images in this category of question which are depicted in Figure 3.

In addition, participants are asked two demographic questions – about their gender ('male', 'female', 'I prefer not to answer this question') and age. At the beginning of the study participants were informed about their task and provided with a contact address to the author of the study.

[2] `https://docs.google.com/forms/d/e/1FAIpQLScGASkwOhnrnzx-OZZ2L3Ewqw3CIkSM8sIJgoifOcgSY01VvA/viewform`.

[3] The original question in Polish is formulated as follows: "Na skali od 1 (nie podoba mi się) do 5 (bardzo mi się podoba), na ile podoba Ci się ten obraz?"

[4] The original question in Polish is formulated as follows: "Jeden z obrazów został stworzony przez sztuczną inteligencję (ang. *artificial intelligence*, AI), a drugi przez człowieka. Który obraz Twoim zdaniem został stworzony przez sztuczną inteligencję (AI)?"

[5] The original question in Polish is formulated as follows: "Czy Twoim zdaniem ten obraz został stworzony przez sztuczną inteligencję (AI)?"

4 Results

Data was collected from 19 individuals (63.2% women) between the age of 18 and 42 (M=27; STD=7.5). On average, accuracy of recognition of AI-generated images reached 61.2%. Three individuals achieved the maximum accuracy equal to 87.5%, and one individual did not recognised correctly any of the images. In terms of the Turing Test we can say that a machine was able to deceive a human judge in 38.2% of cases, on average. In the parallel-paired version of TT AI-generated images were recognised correctly in 73.7% of cases, and in the *viva voce* TT setup accuracy reached 40.4%, on average.

Patterns of responses given by participants that achieve recognition rate above and below the chance level (50%) are depicted in Figure 4. The "below" group scored lower (on average) in terms of accuracy of recognition than the other group in regard to all but one question (C3). This group also rated all images as more aesthetically pleasing than the second group. Interestingly, almost all individuals in the group "above" correctly recognised images B2, B3, and B4 (recognition rate above 90%).

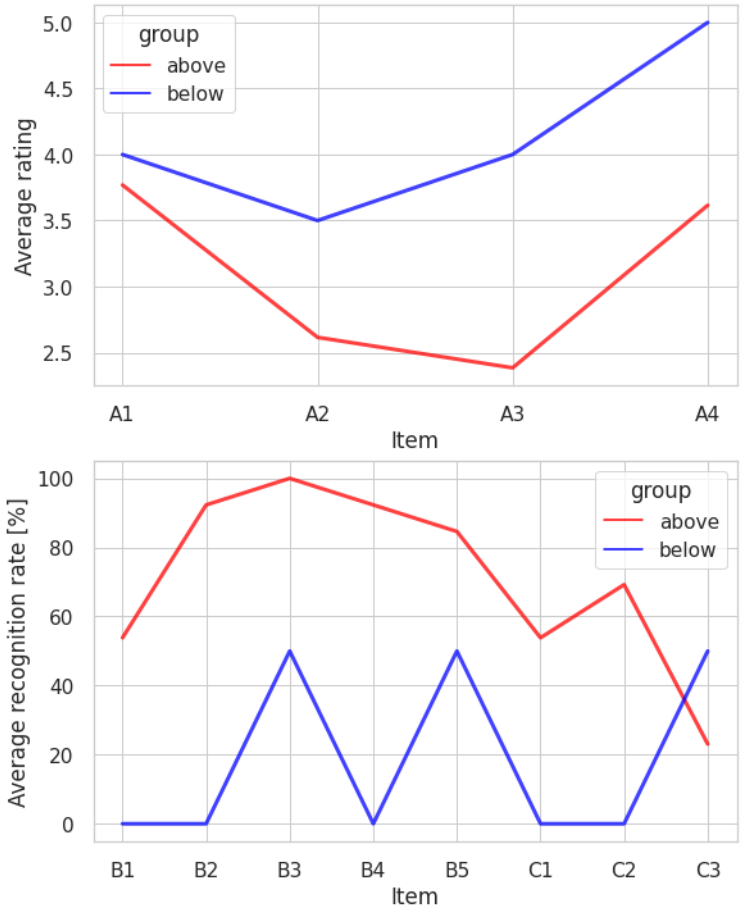

Figure 4: Results of the participants that achieve recognition rate above and below the chance level (50%). Top: Average rating of aesthetic value given by the participants in both groups. Bottom: Recognition rate for each questionnaire item achieved by each group.

4.1 Judgement of Aesthetic Value

First type of questions in the study regards aesthetic value of paintings generated either by AI (questions A1 and A3) or a human artist (questions A2 and A4). Obtained results are summarised in Table 1 and depicted in Figure 5. The highest score was given to the image 1 (A1) – 4 on average (on a scale from 1 to 5). This image was generated by the AI system DALLE. The lowest rating was assigned ex aequo to the image 2 (A2) generated by AI and the image 3 (A3) created by a human.

AI-generated paintings (A1 and A3) achieved the rating of aesthetic value equal to 3.4 (STD = 0.8), and human generated artwork reached the value of 3.2 (STD = 0.8) on average.

The conducted two-sample independent t-test yielded no statistically significant differences in terms of judgement of aesthetic value between AI- and human-generated images at the alpha level 0.05 (t=0.710, p=0.482).

Item	M	STD	Category	M Category	STD Category
A1	4	1	AI	3.4	0.8
A3	2.7	1.3	AI		
A2	2.7	1.2	Human	3.2	0.8
A4	3.6	1.1	Human		

Table 1: Summary of responses of aesthetic value of images. Average rating (M) and standard deviation (STD) is calculated for individual images (A1, A2, A3, A4) and each image category (AI-generated vs. human-generated).

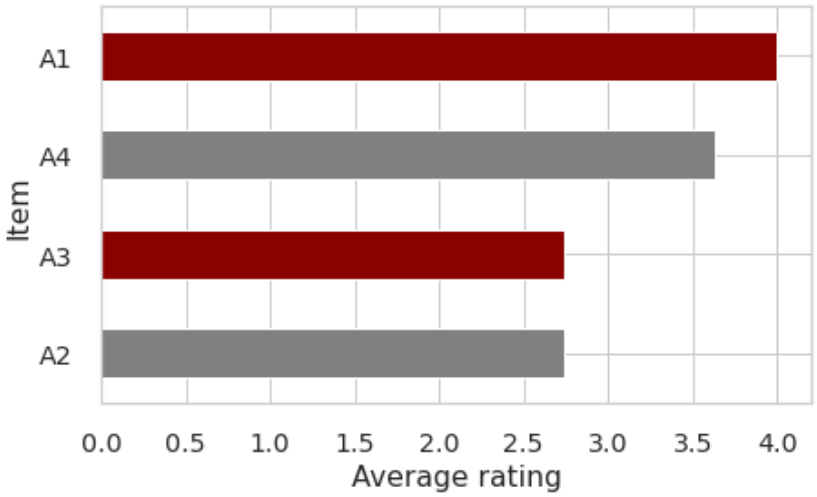

Figure 5: Average rating of aesthetic value assigned to each of the images (A1, A2, A3, A4). AI-generated images are marked by red colour.

4.2 Parallel-paired TT

One-sample t-test, suitable for small sample sizes (less than 30), yielded statistically significant results, i.e. recognition rate above the chance level of 50%, for the following four pairs of images:

- B2 – recognition rate: 79.0% (t=3.013, p=0.008),
- B3 – recognition rate: 94.7% (t=8.5, p<0.001),
- B4 – recognition rate: 84.2%(t=3.980, p<0.001),
- B5 – recognition rate: 73.7% (t=2.282, p=0.035).

Recognition rate for the pair B1, although below the chance level of 50% (36.8%), did not reach a significance level of 0.05 (t=–1.157, p=0.262). Overall recognition rate of 73.7% turns out to be statistically significant (one-sample t-test: t=5.215, p<0.001).

Item	Accuracy	Correct Answer
B1	36.84	Image 1
B2	78.95	Image 1
B3	94.74	Image 2
B4	84.21	Image 1
B4	73.68	Image 1

Table 2: Average recognition rate of AI-generated images in the parallel-paired version of a modified TT.

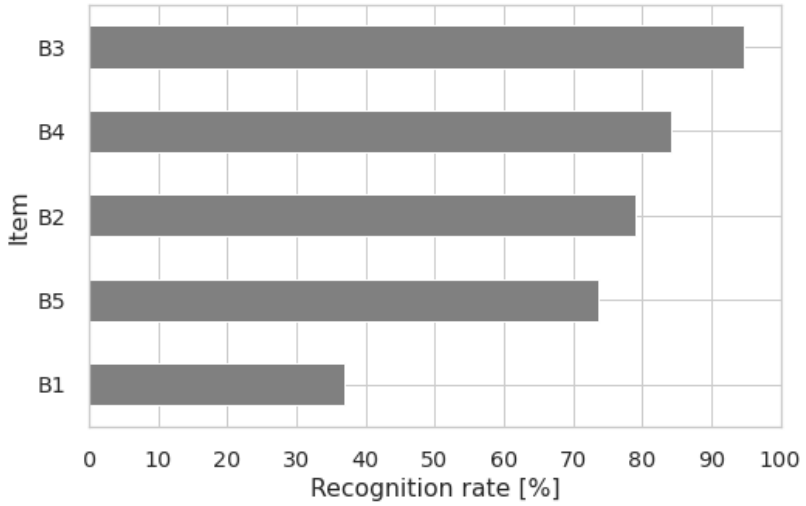

Figure 6: Recognition accuracy achieved by the participants in regard to each pair of images (questionnaire items B1-B5).

4.2.1 Qualitative analysis

In addition to quantitative analysis of results, a qualitative (linguistic) analysis was conducted based on the open-ended responses provided by the participants. In the open-ended form participants were asked to justify their choice of an image that in their opinion was generated by AI. Textual content from these responses was normalised. That is, text was converted to lowercase; then, stop-words (function words) and punctuation marks were removed; finally, text was lemmatised, i.e. words were converted to their dictionary forms. SpaCy library available for Python programming language was used for this purpose. Result of lexical analysis is depicted in Figure 7 as maps of words (so-called word-clouds).

Regarding the first pair of images (B1), participants focused on the shapes of animals painted in the artworks as well as the texture of the images' backgrounds. Participants that chose the correct answer (the image with a dog) referred to "too ideal" shapes of the astronaut dog, however, participants that chose the other image from the pair mention a bit muddled display of the image. The strange combination of a dog and an idea of an astronaut is described as both "chaotic" (and thus ascribed to AI) and "creative" (and therefore claimed to be human-generated). Regarding B2, participants again pointed to "too ideal" shapes of elements as well as the (odd) combination of multiple topics in the image generated AI. Human-generated image was recognised by almost half of the participants in B3, however others mention fewer details painted in the image generated by AI as the reason for choosing the image 2. Image generated by AI in the B4 pair was described as looking "computer-generated" by a few individuals. There two interesting comment opposing each other. One individual said that the image 1 (human-generated) could be created only by a human because semantic knowledge about

politics and Donald Trump is required to create it. On the other hand, someone said that it is a combination of several topics and therefore could be generated by an AI system given adequate keywords in the prompt. Lastly, AI-generated image was again described as an attempt to combine several ill-suited topics (a horse and an astronaut). However, six individuals wrote that they cannot decide and just guess the correct answer.

Figure 7: Maps of common words in open-form responses provided by participants in regard to each of the images – top: B1, B2, B3, respectively; bottom: B4, B5, respectively.

4.3 Viva Voce TT

One-sample t-test yielded one statistically significant result for accuracy of recognition – image C3 with recognition rate equal to 21.1% (t=–3.013, p=0.008). Recognition rates for the other two images (C1 and C2) – 42.1% and 57.9%, respectively – did not reach a significance level of 0.05 (t=–0.678, p=0.506; t=0.678, p=0.506). Overall recognition rate of 40.4% is not significantly different from the chance level of 50% (one-sample t-test: t=–1.472, p=0.147).

Item	Accuracy	Correct Answer
C1	42.11	Human
C2	57.89	AI
C3	21.05	AI

Table 3: Average recognition rate of AI-generated images in the *viva voce* version of a modified TT.

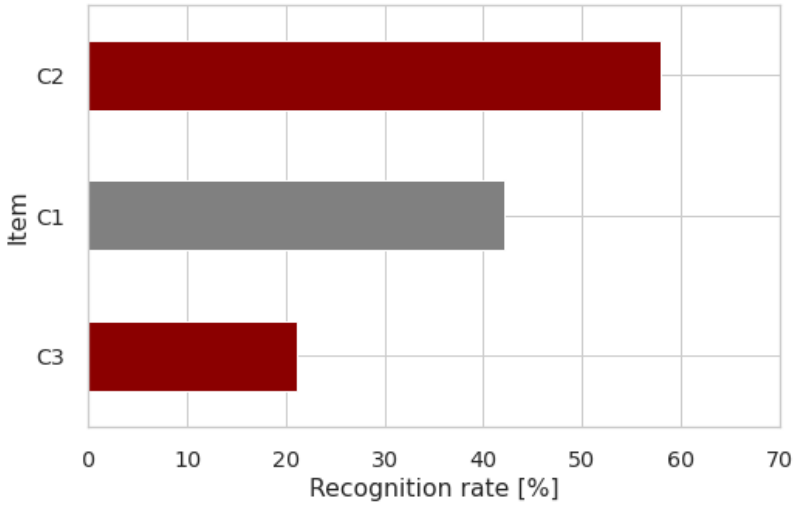

Figure 8: Recognition accuracy achieved by participants in the *viva voce* version of a modified TT. AI-generated images are marked by red colour.

5 Discussion

A machine is said to succeed in the Turing Test if a judge decides wrongly that a machine is a man and a man is a machine at least 30% of the time (Shah, 2010, p. 23). Although the objective of the current study was not to explicitly test whether AI systems could pass TT, it followed the idea of TT. Specifically, the outcomes of an AI system were evaluated against the ones created by a human, instead of the AI system per se as in the original Turing's idea. Images were selected so as to best parallel each other in terms of presented topics/content. Assuming that a machine is less skilled in terms of creating art than humans, one could say participants decided which (in B1-B5 questionnaire items) or whether (C1-C3 items) an AI-generated image falls short of (is less ideal than) a human equivalent or standard. Framing the task in this manner would correspond closely to the Riedl (2014)'s idea of the updated Lovelace Test. Then, one would have to say participants in the study in fact decided which or whether an image satisfies the requirement of generating an image, for example, with an animal in case of B1, a dog in B2, a caricature of Donald Trump in B4, and so on.

In regard to the *viva voce* version of the employed Lovelace-like Turing Test, participants did no better than a random guessing in determining whether a piece of artwork (a painting) was generated by a machine or a human (40.4% accuracy rate). On the other hand, the judges could distinguish between AI-generated and man-made images in the employed parallel-paired TT significantly above the chance level (73.7% accuracy rate).

Specifically, In the parallel-paired version of Turing Test four out of five images were recognised significantly above the chance level. On the other hand, in two out of three cases in the *viva voce* version of the Turing Test recognition accuracy does not differ from a chance level, and in one case (image C3) accuracy of recognition falls significantly below the chance level (to 21.1%). As a result hypothesis 1 was confirmed in the *viva voce* TT, and rejected in the parallel-paired TT.

One may ask what are the reasons behind the discrepancy in results of these two versions of TT. It might be the case that when a judge has two items to evaluate and decide which one is generated by AI, she looks for the "less ideal" one. In this manner the parallel-pair TT could be easier because of this possibility of comparison between the two images (which were on purpose selected so as to present the same topic, e.g. a dog in case of B2, and Donald Trump in case of B4). Even participants in the open-ended form gave answers that they choose this specific image as AI-generated because it is "less ... than the other", or "too ... than the other". It were not specific features about these AI-generated images that made them immediately stood up as created by a machine. There were rather minor details, besides the pair B3 in which half of the participants just recognised the Klimt's

painting. Also, participants wrote a few times that they just guessing the answer because they cannot tell which one is AI-generated. What is more, although participants mentioned several times that the images in pairs B2 and B5 generated by DALLE present a "weird combination" of ideas (e.g., an astronaut riding a horse in cosmos), most of them still answered incorrectly that the image C3 was not generated by AI, for example, which presents a quite odd topic combination as well (i.e., a dog sitting in the bar).

Participants rated AI-generated images as 3.4 and human-generated paintings as 3.2, on average (on a 1 to 5 Likert-like scale). As a result, hypothesis 2 was confirmed – there are no statistically significant differences in terms of aesthetic value of AI-generated and man-made images (independent two-sample t-test: t=0.710, p=0.482). Nonetheless, results indicate differences in judgement of aesthetic value between individual images as ratings averaged over all participants range from 2.7 for images A2 and A3 to 4 in the case of the image A1.

Two main conclusions emerge from the current study that deserve further consideration in particular. First, current state-of-the-art AI systems for image generation (DALLE) have the capacity to create piece of artworks that are judged as aesthetically pleasing as the human-created paintings. Thus, AI systems are able to achieve human-level quality in the area of image generation. Second, the current study indicates that the result of the Turing Test – whether or not a machine could succeed in TT – might depend on the setup of the test (*viva voce* vs. parallel-paired). It would be interesting to further investigate this discrepancy in results in a large sample study and/or with the usage of a different type of stimuli (for example, linguistic samples).

What is more, the current study could be extended further by following the Riedl (2014)' idea of the updated Lovelace Test. In this setup the DALLE system could be provided with a proper prompt containing all the constraints specified by a judge. Then, a human (artist or a lay person) would be provided with the same task of creating a painting, and then the judge would compare the two and tell which one satisfies the requirements. Again, a few rounds of testing would be conducted with not only different topics of images to be generated but also different judges. The result then would be the proportion of tasks an AI system generates an image that satisfies the requirements orderred by a judge.

References

M. A. Boden. The turing test and artistic creativity. *Kybernetes: The International Journal of Systems & Cybernetics*, 39(3):409–413, 2010.

S. Bringsjord, P. Bello, and D. Ferrucci. Creativity, the turing test, and the (better) lovelace test. *Minds and Machines*, 11:3–27, 2001.

F. Carnovalini and A. Rodà. Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence*, 3:14, 2020.

R. Chamberlain, C. Mullin, B. Scheerlinck, and J. Wagemans. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, 12(2): 177–192, 2018.

R. M. French. Dusting off the turing test. *Science*, 336(6078):164–165, 2012.

J.-W. Hong and N. M. Curran. Artificial intelligence, artists, and art: attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–16, 2019.

N. Köbis and L. D. Mossink. Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate ai-generated from human-written poetry. *Computers in human behavior*, 114:106553, 2021.

N. C. Köbis, B. Doležalová, and I. Soraperra. Fooled twice: People cannot detect deepfakes but think they can. *Iscience*, 24(11):103364, 2021.

A. Pease and S. Colton. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *Proceedings of the AISB symposium on AI and Philosophy*, volume 39, 2011.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

M. Ragot, N. Martin, and S. Cojean. Ai-generated vs. human artworks. a perception bias towards artificial intelligence? In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–10, 2020.

M. O. Riedl. The lovelace 2.0 test of artificial creativity and intelligence. *arXiv preprint arXiv:1410.6142*, 2014.

J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980. doi: 10.1017/S0140525X00005756.

H. Shah. *Deception-detection and machine intelligence in practical Turing tests*. PhD thesis, University of Reading, 2010.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.