Contemporary Creativity: The many Faces of AI Art

A research project on the creative potential of Dream-OOD AI-generated images through the lens of Boden's Creativity Framework using an Elo-based rating system

Jari de Keijzer[1]
Supervisor: Dr. A. Lukina[1]
[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements For the
Bachelor of Computer Science and Engineering
July 26, 2024

Name of the student:	Jari de Keijzer
Final project course:	CSE3000 Research Project
Thesis committee:	Dr. A. Lukina & Petr Kellnhofer

Abstract

This research project investigates the creative potential of AI-generated images, specifically those produced by the Dream-OOD diffusion model, through the lens of Margaret Boden's creativity framework. We conducted a user study employing an Elo-based rating system to assess the novelty, surprisingness, and value of Dream-OOD generated images, comparing them to real-world IMAGENET images. The results show that Dream-OOD excels in generating images perceived as novel and surprising while aligning with Boden's concept of exploratory creativity. However, AI-generated images often fall short in terms of value compared to real images. This study serves as a proof of concept for using Boden's framework together with an Elo based ranking system, to evaluate AI creativity, highlighting the potential of Dream-OOD in generating novel and surprising content while acknowledging its limitations in replicating the aesthetic and emotional depth of human-created art.

1 Introduction: creativity or imitation

The increasing prevalence of AI tools has reignited the debate on creativity: What is human creativity and can AI replicate it? If so, to what extent? If not, what sets them apart? This research project delves into these questions by examining the creative potential of the Dream-OOD model, a novel diffusion model that generates images with seemingly creative attributes. We aim to assess whether AI-generated images can be considered creative using Margaret Boden's established framework of creativity by ranking them using an Elo rating system.

Boden[1] defines creativity as generating novel, surprising, and valuable ideas. These ideas can be novel to the individual (P-creativity) or novel to the world (H-creativity). Boden argues that AI should focus on P-creativity, as achieving this could lead to H-creativity. She also identifies three types of creativity: combinational, exploratory, and transformational. These distinctions can be used to analyse the output of AI models.

In this research, we focus on Dream-OOD, a diffusion model that generates out-of-distribution (OOD) images. These images are often unusual and unexpected, suggesting potential creativity (see Figure 1 for reference). We investigate whether Dream-OOD exhibits creativity according to Boden's framework.

To evaluate the creativity of Dream-OOD's output, a user study was conducted where participants rated the novelty, surprisingness, and value of AI-generated images compared to real-world images. An Elo rating system was used to rank the images based on these criteria.

The results show that Dream-OOD excels in generating images perceived as novel and surprising, aligning with Boden's concept of exploratory creativity. However, AI-generated images often fall short in terms of value compared to real-world images. This suggests that while AI can generate novel and surprising content, it can struggle to replicate the aesthetic and emotional depth of human-created imagery.

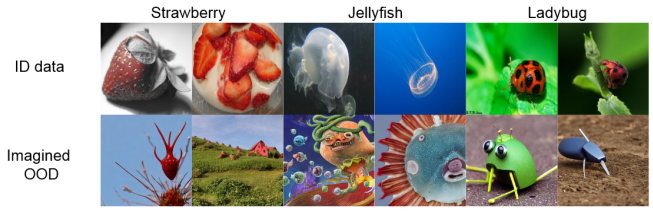


Figure 1: IMAGNET ID training data vs. imagined OOD samples generated by Dream-OOD [2]

This paper is structured as follows. Section 2 provides background information on Boden's creativity framework, the `Dream-OOD` model and the Elo rating system. Section 3 outlines the research question and sub-questions that this paper answers. Section 4 details the methodology, including the design of a user study and the Elo rating system. Section 5 presents the results of the user study, analysing the Elo scores and the ranking of the images. Section 6 summarises our main findings and contributions. Section 7 discusses the implications of our findings, addresses the limitations of the study, and suggests directions for future research. Finally, Section 8 concludes the paper by going into the ethics of this research project.

In conclusion, we find that Dream-OOD is capable of generating novel and surprising images, aligning with Boden's concept of exploratory creativity. However, AI-generated images often lack the value associated with real-world images, indicating a limitation in replicating the aesthetic and emotional depth found in human-created art.

2 Background

This research paper mainly builds on three papers. The first gives a framework for creativity that in turn can be used to evaluate the creativity of an AI model, the second is a new diffusion model that generates images that appear to have some creative features, the third is the original text that describes how to rank chess players, also known as the Elo rating system. This paper is a proof of concept of a method for evaluating creativity in Artificial Intelligence models. It does this by applying the creativity framework and the ranking system to the new diffusion model.

2.1 A framework for creativity

In the 1990's Margaret Boden wrote a book [3] and a paper [1] on creativity and artificial intelligence. Her book focusses mainly on the questions surrounding creativity, where she uses computational examples to help grasp what creativity is. Before diving into these examples, she acknowledges that many people would intrinsically say, or feel, that creativity is something inherently human. Computers would only be able to create and do what they are programmed to do. She mentions that Lady Ada Lovelace is the first to make this argument in some way; Boden agrees with the statement itself, just not what it implies, that just following programming

voids a programme of all creativity. So, to counter this, Boden poses four questions, which she calls Lovelace questions.

1. Can computational ideas help us to understand how human creativity is possible?
2. Can computers (now or in the future) ever do things that at least appear to be creative?
3. Can a computer ever appear to recognise creativity?
4. Can computers themselves ever really be creative (as opposed to merely producing apparently creative performance whose originality is wholly due to the human programmer)?

The main focus of her book is on the first Lovelace question; however, the answer to the first three questions is yes, according to Boden. She states that the answers to the first three questions can be scientific, and that these question are really closely related to each other. The fourth question is of a completely different type, it is a philosophical one. In the rest of her book, she builds a framework of creativity [4] which this paper will use.

Boden describes creativity as something that is **novel**, **surprising**, and **valuable**. Where novel can either mean something that has never been seen in the world before, which she calls H(istorical)-creativity; Or it could mean something which has never been seen by the person or computer/programme creating it, but not necessarily new in the world; for instance, a child that makes up a joke that they have not heard of before, making it personally new, but the joke has been made by someone somewhere before. She calls this P(ersonal)-creativity. She focusses on the latter, since most people can still find something to be creative, even though it has been done before in history. She further explains that these three features of creativity are quite a personal experience, where one person might feel something to be creative when another does not. This is also the reason why the fourth Lovelace question is a philosophical one; it does not have a clear answer.

To further define the problem space, she divides creativity into three types: **combinational**, **exploratory**, and **transformational** creativity. Respectively, they mean novel (improbable) combinations of familiar ideas, the generation of novel ideas by the exploration of a structured conceptual space, and the generation of novel ideas by transforming some dimension of the space so that new structures can be generated. In her paper, she concludes that some creative ideas have been created by exploratory or combinational procedures, but that transformational creativity was just beginning and is rather difficult because of two bottlenecks:

- domain expertise, which is required to map the conceptual space that is to be explored and then also transformed
- valuation of results, which is even more difficult for transformational procedures since the results would not be part of the input conceptual space.

To have transformational creativity a programme should be able to transform and adapt by using it's domain expertise and evaluating its result without further inputs of a programme. This would create the distinction between appearing to be creative and actually being creative. Back then no such model existed; maybe the next model has a chance.

2.2 Dream-OOD: a potentially creative framework

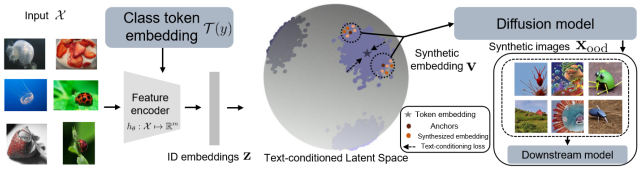


Figure 2: Illustration of their proposed outlier imagination framework `Dream-OOD` [2]

In the paper 'Dream the Impossible' [2], a new model called `Dream-odd` is proposed (Figure 2), "which enables imagining photo-realistic outliers (Figure 1) by way of diffusion models, provided with only the in-distribution (ID) data and classes."

The researchers used 100 IMAGENET [5] classes and their images as training data to learn "a text-conditioned latent space based on ID data, and then sample outliers in the low-likelihood region via the latent, which can be decoded into images by the diffusion model." This means they train an image classifier that produces image embeddings that have a high probability of being aligned with a class token embedding of one of the 100 ImageNet classes and the other way around. From this they learn the text-conditioned latent space, which can be represented as clusters around the token embeddings (Figure 3). A fixed text encoder of the diffusion model is used so that only the image feature encoder is trained.

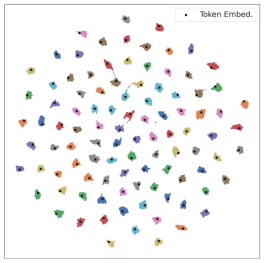


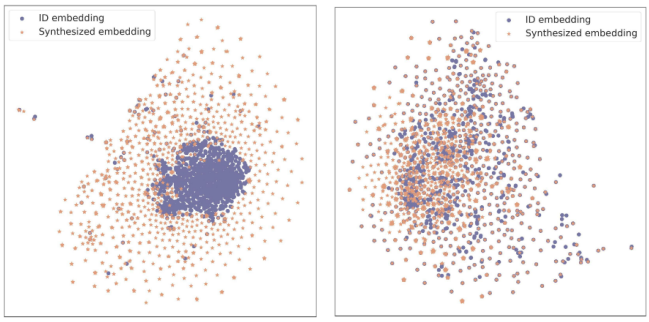
Figure 3: TSNE visualization of learned feature embeddings [2]

In this latent space they create new embeddings by moving away from the centroid of these clusters and use those new embeddings to generate out-of-distribution (OOD) images (Figure 4a). They also generate new ID images (Figure 4b).

The goal of this research is to generate informative outliers that lie on the border of ID data, and use those outliers to generate images that can be used to further advance the research on OOD detection. Their insight is that OOD detection is hard since it is hard for humans to come up with a method to detect outputs that we cannot envision ourselves. OOD images are often rather strange and far of from real-world images. Another way to look at this is that this model generates rather creative images that could be novel, surprising, and valuable. This hypothesis will be tested in this paper.

2.3 Elo rating system

Elo's rating system [6] is best known for its usage on rating the performance of chess players. The main idea of the rat-



(a) Synthesised outliers (b) Synthesised inliers

Figure 4: TSNE visualization of IMAGENET ID embeddings (purple) and the synthesized embeddings (orange), for class "hen" in IMAGENET [2]

ing is to rank players without having to let them play against all other players. The rating is an estimation of the players strength and follows a logistic/normal distribution, both work in practice.

A player's performance is estimated based on wins, losses, and draws against other players. The **expected score** of a player (E) is their probability of winning plus half their probability of drawing. The expected score of player A (E_A) against player B (E_B) is calculated using their current ratings (R_A and R_B) as follows:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}} \tag{2}$$

Note that $E_A + E_B = 1$. So when the players have the same Elo rating, they have an expected win chance of 0.5 each. The originally expected score factor of 400 was arbitrarily chosen such that for two players that differ 200 points, the stronger player has an expected win chance of 0.75.

After a game, the player's rating is updated based on the difference between their actual score (S_A) and their expected score (E_A). The score for winning is 1.0, losing 0,0 and ending in a draw is 0.5. The update formula is as follows:

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{3}$$

Here R_A & R'_A are the current and updated Elo ratings for player A. K is the K-factor, a parameter that determines the maximum possible adjustment per game. In chess matches of stronger players get a lower K factor to defer from update scores too fast. This rating system will be used to rank the creativity of images.

3 Problem definition

As described in the introduction, there has always been a question whether AI can show creativity or merely imitate human creativity. Research [7] has been done to find a link between the creativity of LLMs and their temperature parameter (introduces randomness). However, the relationship between that parameter and creativity remains unclear. Since we have not found a clear parameter to tweak creativity yet, we still need to be able to find and define it in models individually. Hence, there is a need to implement a method to test AI models on their creativity; this paper introduces such a method.

This paper focusses on applying Boden's framework of creativity to the `Dream-OOD` model using an Elo rating system to answer the following research question.

3.1 Research Question

Can Boden's framework of creativity be used to determine whether `Dream-OOD` is creative?

SQ1. What type of creativity could `Dream-OOD` possibly portrait? (combinational, exploratory, transformational)

SQ2. Using images generated by the `Dream-OOD` model, can we rank their creativity along each of Boden's dimensions (novel, surprising, valuable).

SQ3. Can we say that the `Dream-OOD` model appears to be creative?

SQ4. Can we describe what features in the outputs make it appear creative?

SQ5. Can we conclude that the `Dream-OOD` model is creative? Philosophically differentiation between appearing creative and being creative.

4 Methodology

The following subsections will go over the methodological choices that were made to answer each individual research sub-question. They will go over the potential creativity type of the `Dream-OOD` model, the ranking mechanism used to rank the creativity of images, the user study that has been designed to score the creativity of `Dream-OOD`, what features influence this creativity, and if this means `Dream-OOD` is creative or just appears so.

4.1 Type of creativity

As mentioned before, Boden defines three types of creativity [3], `combinational`, `exploratory`, and `transformational`. These three types themselves however, besides having a clear description, do not have an exact definition. Therefore answering the question of which type of creativity could potentially be found in `Dream-OOD` will not have an exact answer either. To get as close as possible to an exact answer, this section will analyse the algorithm that synthesises the `OOD` & `ID` embeddings (shown in Figure 4) which in turn are used to generate the `OOD` & `ID` `Dream-OOD` images, and compare this algorithm with Boden's description of each type of creativity.

To start with `combinational` creativity, which Boden defines as the generation of something novel through the unfamiliar combinations of familiar ideas. What's important here, is that these combinations cannot merely be random, they need to based on some meaningful connection. `Dream-OOD` doesn't have a notation of certain styles, objects, aesthetics that it could combine, so it cannot create meaningful combinations. It also does not combine multiple images into new

images, there is no combination of any kind. Therefore, it shows no `combinational` creativity.

The other two types of creativity revolve around the concept of conceptual spaces, which Boden defines as structured styles of thought (for humans). These spaces have inherent structures and constraints that govern the possibilities in them. These structures and constraints find their origins in culture, social settings, traditions, trends, styles, genres, paradigms in science, mathematics, or any other constraint one can think of. `exploratory` creativity would, as the name suggests, explore within the limits of these conceptual spaces, and `transformational` creativity would transform the conceptual space, moving outside its bounds by altering its form.

Now, the boundary of the conceptual spaces is what makes it interesting. Since these are based on societal values and norms, they can change over time. Maybe an art style per se does not change, but the way we perceive that art does change. Since these boundaries aren't exact and change over time anyways, there isn't an exact point when exploring becomes transforming a conceptual space.

Now for the `Dream-OOD` model, they use a Gaussian function to synthesise new embeddings on the border of a classifier. Although this synthesis follows a certain function, it still follows a random probability function. Considering that these points will still try to be close to a certain class centroid, this can be considered `exploratory` creativity. The model explores the learned text-conditioned latent space. To be considered `transformational` creativity, the model would have to move away further from the class centroids, which the researchers tried, but resulted in non-valuable images. As Boden stated, the difficult part in `transformational` creativity is the valuation of the results, in other words, how the model moves away from a certain conceptual space, but remains valuable. To be considered `transformational` creativity, the model would have to be able to evaluate these options by itself, even when moving further outside of a conceptual space, for now, this is not something `Dream-OOD` is capable of doing. So we will consider `Dream-OOD` to be of the `exploratory` type.

4.2 Ranking creativity

In terms of how the Elo rating system is used to rate the chess player's performance, we used it to rate the creativity of images. In our situation, the images were the 'players' of a game. In each game, an image could win or lose in one of the creativity features of Boden (`novelty`, `surprise`, `value`). Depending on those wins and losses, an updated score can be calculated using the formulas described in section 2.3. The combination of `novelty`, `surprise`, and `value` wins and losses will be made by dividing their total by three. So winning in `novelty` and `surprise` but losing in the `value` category (like in Figure 5), will result in a score of 2/3 and 1/3 for the picture pair.

4.3 User study: voting on random pairs of images

To test whether `Dream-OOD` could appear creative, a user-study has been set up in the form of a voting website (creativity-in-ai.ewi.tudelft.nl) where participants could vote on randomised pairs of images. After going through an opening statement (Appendix A) and an explanation page (Appendix B), the participants were shown 30 random pairs of images. For each pair, the participant could select which of the two images they found the most novel, surprising, or valuable. After which, they could submit their selection, effectively sending in three votes per pair of images. They would do this for at least 30 votes (pairs) and could increase their number of votes in increments of 10 after the first 30. Figure 5 shows the voting interface that participants would see; in this situation, the participant has chosen the right image (tiger) to be more Novel and Surprising and the left image to be the most valuable. Each participant voted with their own preferences following their intuition.

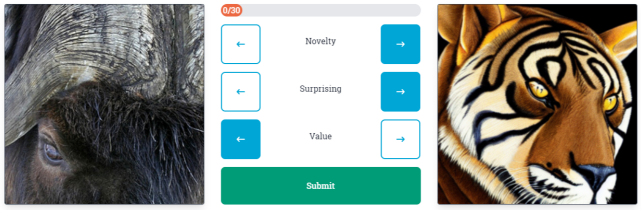


Figure 5: Voting interface for the user-study participants

Due to the expected number of participants and votes, 60 images were selected to be ranked. These images were selected from a single `IMAGENET` class, specifically the `0x` class. This class has been chosen for its aesthetic value. These decisions will be reflected upon in the section 7.3. The images themselves were randomly picked by using a random integer generator and selecting the i-th image of the different image types. Pre-processing included cutting the images to one image size (256px by 256px), taking the centre portion of the image if an image was not square and was larger than the selected image size. All images were given a random integer from 1 to 60 to make the image type unidentifiable for the participants.

The following division of images has been made:

- 20 `Dream-OOD` Out-Of-Distribution (`OOD`) generated images, that had a high deviation from the class centroid
- 20 `Dream-OOD` In-Distribution (`ID`) generated images, that had a low deviation from the class centroid.
- 20 `IMAGENET` real-world images.

4.4 Finding features of creativity

To find some features that give the images a higher Elo score, we will perform some visual inspections on the rankings. Even though these creativity features are quite personal, we still hope to be able to observe some features that stand out that make these images get a high or low score in their different creativity features.

4.5 Being vs appearing creative

After finding whether Dream-OOD can appear creative for one of the creativity types by using the ranking described before, we will briefly go over what this means for actually being creative. We will go over when we could define something to be creative when it appears to be, and if this is possible altogether, and whether this is important at all.

5 Results

Over a time frame of two weeks, the user study accumulated a total of 4222 votes from 151 participants. As mentioned before all users were required to vote for a minimum of 30 times and could increase their number of votes in steps of 10. The tables and plots in this section show values for one of three different scenarios (filters), although there were images with a different position in the ranking (due to some limited votes per pair), there were no significant differences between the scenario's on which we based our conclusions. This is shown by high agreement in rankings between the different scenarios shown in Table 1.

Table 2 shows the three scenario's (filters): `all votes`, `30 minimum`, and `30 exactly`. Where `30 minimum` also included the plus 10 increments of votes above 30 votes per participant (described in section 4.3). We have chosen to show scenario `30 exactly` due to the personal nature of creativity. For this scenario, each participant had the same amount of input, exactly 30 votes. The rest of the results will be posted together with this paper.

Table 2: Vote filtering scenarios

Scenario	Participants	Votes	Difference
all votes	151	4222	-
30 minimal	112	3890	332
30 exactly	112	3360	530

5.1 Number of wins per type and feature

Table 3 shows the number of winning votes per creativity feature over the different image types. It shows that the number of wins in `Novelty` and `Surprise` are fairly similar to each other. Besides that, the `OOD` & `ID` images win the most on the `Novelty` and `Surprise` features, and the `IMAGENET` images win the most on the `Value` feature. Lastly, the `OOD` & `ID` images also have more winning votes in total, compared to `IMAGENET` images.

Table 3: Win table

Category Image Type	Novelty	Surprise	Value	Total
ID	1326	1327	1023	3676
IMAGENET	626	692	1332	2650
OOD	1408	1341	1005	3754

Table 4 shows the Chi-squared test significance between the different image types. It shows significant p-values for all image types. It confirms that there are significant differences in the number of votes on the different image types and creativity features. In addition, it shows significant differences within each image type comparing the different creativity features.

Table 4: Chi-squared test win significance table

Image Type	Chi-squared statistic	DoF	P-value
OVERALL	468.947	4	**3.479e-100**
OOD	74.532	2	**6.539e-17**
ID	50.116	2	**1.311e-11**
IMAGENET	344.299	2	**1.724e-75**

Table 5 shows the standardised residuals of the chi-squared `OVERALL` test. It confirms that `IMAGENET` has a significantly higher score in `Value`, and a lower score in `Novelty` and `Surprise`. The opposite holds for `OOD` & `ID` Images.

Table 5: Standardised residuals after chi-squared on wins

Image Type	Category	Z-score
OOD	Novelty	4.429
	Surprise	2.535
	Value	-6.964
ID	Novelty	2.876
	Surprise	2.904
	Value	-5.780
IMAGENET	Novelty	-8.658
	Value	15.096
	Surprise	-6.438

5.2 Agreement between creativity features

Due to some participants privately responding that it was difficult to distinguish between `Novelty` and `Surprise`, some calculations were made to determine if this was true for most participants. We looked at the agreement between the three different creativity features.

In figure 6 the four different combinations of the three creativity features were tested against the three image types. It shows that the votes on `Novelty` and `Surprise` (`N&S`) have a strong overlap, so in 39.8 percent of the votes people who voted for a certain image to win on the `Novelty` feature would also select that image to win for the `Surprise` feature, and vice versa. This percentage is even higher within the `OOD` & `ID` type. In addition, having an image win on all three creativity features is more common overall than winning in both `Novelty` & `Value` (`N&V`) or in both `Surprise` and `Value` (`S&V`) . Only for the `IMAGENET` type, it is more common to win in all three categories than in any of the three sub-combinations.

Table 1: Correlation between scenario rankings using Kendall's tau, Spearman's rho, Somers' Dxy

Rank Type	Comparison	Kendall's tau statistic	Kendall's tau p-value	Spearman's rho statistic	Spearman's rho p-value	Somer's Dxy statistic	Somer's Dxy p-value
Combined Rank	all votes vs 30 minimal	0.937	**5.006e-26**	0.991	**2.701e-52**	0.938	**0.000e+00**
	all votes vs 30 exactly	0.912	**1.012e-24**	0.985	**2.491e-46**	0.913	**0.000e+00**
	30 minimal vs 30 exactly	0.936	**5.369e-26**	0.992	**2.324e-53**	0.935	**0.000e+00**
Novelty Rank	all votes vs 30 minimal	0.964	**1.417e-27**	0.996	**3.646e-63**	0.964	**0.000e+00**
	all votes vs 30 exactly	0.953	**6.091e-27**	0.994	**1.044e-57**	0.952	**0.000e+00**
	30 minimal vs 30 exactly	0.957	**4.001e-27**	0.996	**4.571e-62**	0.956	**0.000e+00**
Surprise Rank	all votes vs 30 minimal	0.950	**9.173e-27**	0.993	**3.092e-55**	0.951	**0.000e+00**
	all votes vs 30 exactly	0.922	**2.551e-25**	0.990	**2.583e-51**	0.923	**0.000e+00**
	30 minimal vs 30 exactly	0.939	**3.164e-26**	0.993	**1.965e-55**	0.939	**0.000e+00**
Value Rank	all votes vs 30 minimal	0.926	**1.599e-25**	0.989	**4.290e-50**	0.925	**0.000e+00**
	all votes vs 30 exactly	0.890	**1.150e-23**	0.980	**1.582e-42**	0.889	**0.000e+00**
	30 minimal vs 30 exactly	0.910	**1.326e-24**	0.986	**6.634e-47**	0.909	**0.000e+00**

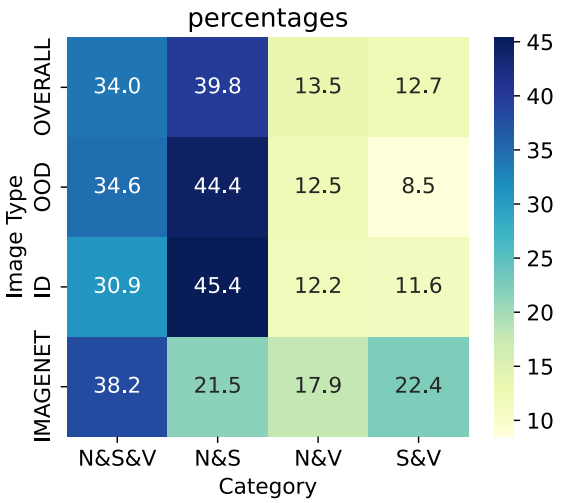


Figure 6: Agreement between creativity features

Table 6 shows the chi-squared test significance between the agreement of votes for each of the image types. It clearly shows by the low p-values that there are significant differences between the agreement of voting categories.

Table 6: Chi-squared test agreement table

Image Type	Chi-squared statistic	DoF	P-value
OVERALL	173.923	6	**6.619e-35**
OOD	481.588	3	**4.663e-104**
ID	411.432	3	**7.393e-89**
IMAGENET	71.289	3	**2.261e-15**

Table 7 shows the standardised residuals of the chi-square test for the image types and combinations of agreement. This shows that the `IMAGENET` category deviates the most in the `N&S` and the `S&V` category.

Table 7: Standardised residuals agreement table

Image Type	Category	Z-score
OOD	N&S&V	0.421
	N&S	2.669
	N&V	-1.044
	S&V	-4.341
ID	N&S&V	-1.911
	N&S	3.166
	N&V	-1.321
	S&V	-1.117
IMAGENET	N&S&V	1.981
	N&S	-7.846
	N&V	3.180
	S&V	7.377

5.3 Elo scores calculations

Because of the high significant agreement differences over the creativity features, which meant that they were not entirely independent, we decided to calculate seven different Elo scores. Three of which are Elo scores when only a single creativity feature is taken into account. Another three, are the Elo scores when taking two of the three creativity features. The seventh score is the combined score that takes into account all three creativity features. The combined score uses the intended creativity framework that Boden proposed.

For the Elo calculation we used a start rating of 1500 and a K-factor of 32. These were the initial values Elo used as well. The tweaks that many chess associations have implemented would not apply to our case, so we used this initial set-up.

Figure 7 shows the different mean of the seven different Elo scores over time. The biggest plot shows the combined scores. It shows that over time the mean score of `OOD` & `ID` images keep on rising and the mean score of `IMAGENET` images keep on falling. Also, for the `Value` feature it shows a higher mean for `IMAGENET` images and lower for the rest. In addition, the `SV Score` shows almost no difference in mean value over the different image types, whereas the `NS Score` shows a very large difference between the `OOD` & `ID` images and the `IMAGENET` images. The agreement we calculated before in section 5.2, have some impact on this.

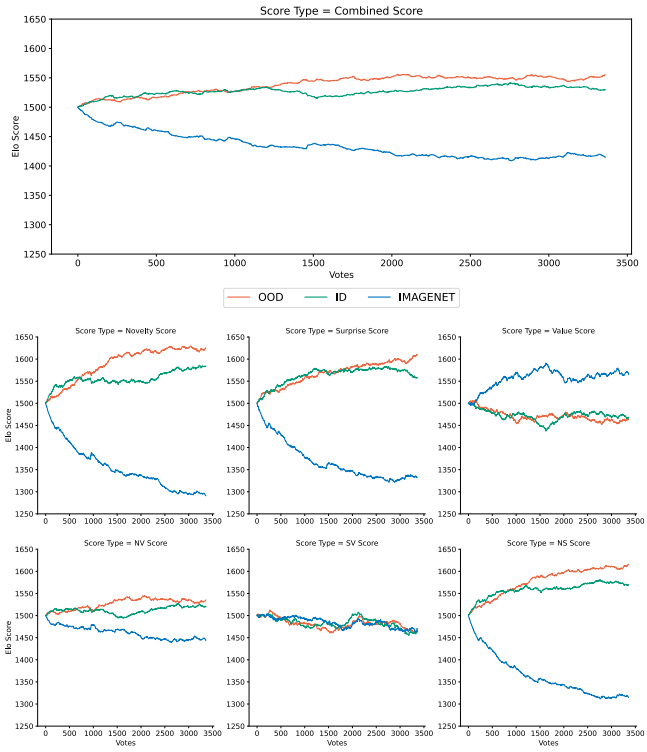


Figure 7: Mean Elo scores over time

Figure 8 shows the KDE distribution of the different score types that were calculated, with a fitted logistic distribution in red. What is noteworthy here, is that scores of the separate creativity features are more spread out, as would be expected when only taking one feature, but when combined the scores have less variance. In addition, the `SV` score shows a very high peak, which means the scores are very much centred around the starting score of 1500, showing that the `Surprise` and `Value` features seem to be each others opposites in overall scores.

Table 8 shows the p-values of three different logistic distribution test run on the seven different score types after they were calculated. The test show no significant p-values, which means that we cannot reject the null hypothesis of the Elo score distribution being logistically distributed. Exactly as we had seen in Figure 8, all scores follow a logistic distribution, exactly how Elo intended the scores to be.

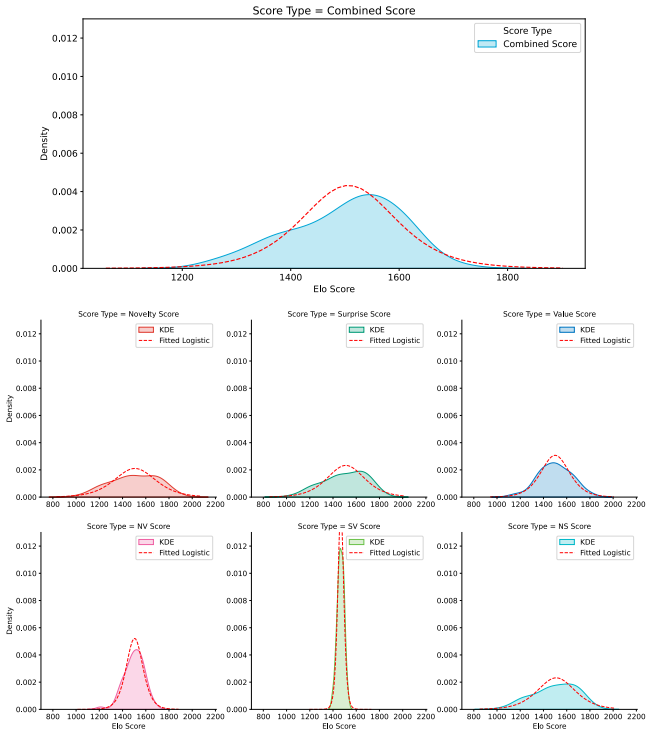


Figure 8: Distribution of ELO Scores with Fitted Logistic Distributions

Table 8: Elo calculation distribution fit tests

Distribution test	Score Type	p-values
Kolmogorov-Smirnov Logistic Test (p-value)	Combined Score	0.743
	Novelty Score	0.582
	Surprise Score	0.348
	Value Score	0.952
	NV Score	0.823
	SV Score	0.804
	NS Score	0.734
Anderson-Darling Logistic Test (statistic)	Combined Score	0.686
	Novelty Score	0.628
	Surprise Score	0.765
	Value Score	0.247
	NV Score	0.273
	NS Score	0.714
	SV Score	0.292
Cramér-von Mises Logistic Test (p-value)	Combined Score	0.636
	Novelty Score	0.673
	Surprise Score	0.577
	Value Score	0.946
	NV Score	0.959
	NS Score	0.658
	SV Score	0.883

5.4 Elo score significance

Figure 9 shows box plots of the seven different Elo scores between the different image types. Most of the OOD & ID scores are relatively higher than the IMAGENET scores. Within the IMAGENET type, the different score's show a larger variance. What is also noteworthy is that the singular scores (Novelty, Surprise, and Value) show a larger spread in general, but when the scores types are combined (except the NS score) the spread is less and the scores seem to settle.

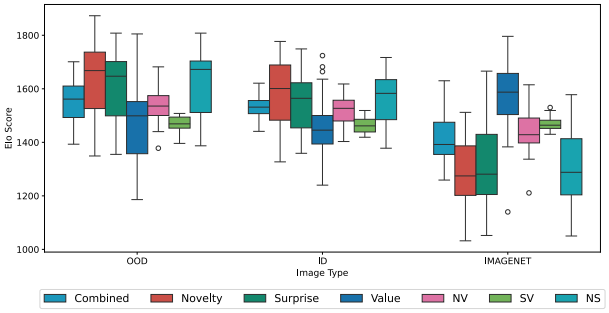


Figure 9: ELO Scores by Image Type and Score Category

Table 9 shows the p-values of the Kruskal-Wallis (KW) statistic together with the Dunn's and Conover-Iman post hoc tests. KW shows if there are significant differences in the Elo scores. For an alpha of 0.05, only the KW test does not conclude a significant difference in the SV score.

The post hoc tests show if there are significant differences between the individual image types. Both post hoc tests confirm that there is a significant difference in Elo scores between OOD images compared to IMAGENET images and ID images compared to IMAGENET images. Only the SV score does not show significant differences. The test also shows that there are no significant differences in Elo scores between the OOD images and the ID images.

5.5 The ranking: visual inspection

Figure 10 shows that the overall rankings of OOD & ID images are lower than the IMAGENET rankings, with the exception of the Value and SV rank. A higher Elo results in a lower ranking in this plot, where position 1 in the ranking has the highest Elo score and position 60 the lowest Elo score.

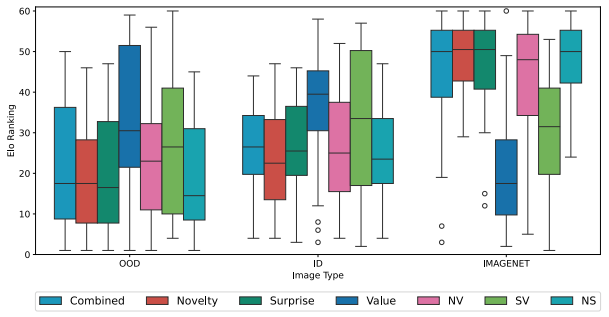


Figure 10: ELO Rankings by Image Type and Score Category

Figure 11 shows the visualised ranking of the Combined Rank. It shows that the high ranking OOD & ID images often are very novel and surprising, in the sense that they are quite obviously AI generated, but also not trying to imitate something in the real world, they seem creative in their own way. The lower ranked OOD & ID images seem to show novel and surprising things which we as humans cannot relate to, too odd or strange looking for us to find it valuable. The higher ranked IMAGENET images seem to have a certain aesthetic value, with things like sharp zoomed in images, a good looking background, or nice colours. The lower ranked IMAGENET images seem to just be very blurry or have no focus on a certain object. These observations are personal observations, but still quite general.

6 Conclusion

Using the methodology described in chapter 4 and combining the different results, we can formulate a number of conclusions:

1. The results for three scenario's were calculated which showed no significant differences in rankings. The rest of the results only showed the 30 exactly scenario.
2. Wins:
   - OOD & ID images win most on Novelty and Surprise and are similar in number of wins.
   - IMAGENET images win the most on Value.
   - These differences were shown to be significant.
3. Agreement:
   - There is significant agreement between the Novelty & Surprise feature, even more so for the OOD & ID images.
   - This can be explained by anecdotal evidence that the participants found it hard to distinguish between those two features.
4. Elo calculations:
   - The mean Elo scores over time followed the same conclusions as for the number of wins per image type. OOD & ID higher scores and IMAGENET lower scores.
   - The mean Elo scores of SV show almost no differences over time. These scores must be each others opposites, overall.
   - All scores followed a logistic distribution.
5. Elo scores
   - The box plots showed OOD & ID images to have higher scores than IMAGENET images.
   - Only the Value scores showed the opposite.
   - All of these differences in scores were shown to be significant for the OOD vs IMAGENET and ID vs IMAGENET, except for the SV Score
   - There were no significant differences between OOD & ID images.

Table 9: Kruskal-Wallis, Dunn's & Conover-Iman on scores

Score Type	Kruskal-Wallis statistic	Kruskal-Wallis p-value	Dunn's Test (p-value) OOD vs. IMAGENET	Dunn's Test (p-value) ID vs. IMAGENET	Dunn's Test (p-value) OOD vs. ID	Conover-Iman Test (p-value) OOD vs. IMAGENET	Conover-Iman Test (p-value) ID vs. IMAGENET	Conover-Iman Test (p-value) OOD vs. ID
Combined Score	18.512	**9.555e-05**	**1.183e-04**	**0.005**	1.000	**2.700e-05**	**0.001**	0.785
Novelty Score	31.628	**1.355e-07**	**6.862e-07**	**2.084e-05**	1.000	**1.590e-09**	**6.798e-08**	0.994
Surprise Score	23.803	**6.781e-06**	**9.570e-06**	**0.001**	0.850	**5.662e-07**	**8.204e-05**	0.533
Value Score	8.373	**0.015**	**0.038**	**0.035**	1.000	**0.031**	**0.030**	1.000
NV Score	11.606	**0.003**	**0.005**	**0.020**	1.000	**0.003**	**0.013**	1.000
SV Score	0.775	0.679	1.000	1.000	1.000	1.000	1.000	1.000
NS Score	28.803	**5.565e-07**	**1.622e-06**	**9.167e-05**	1.000	**1.483e-08**	**1.197e-06**	0.756

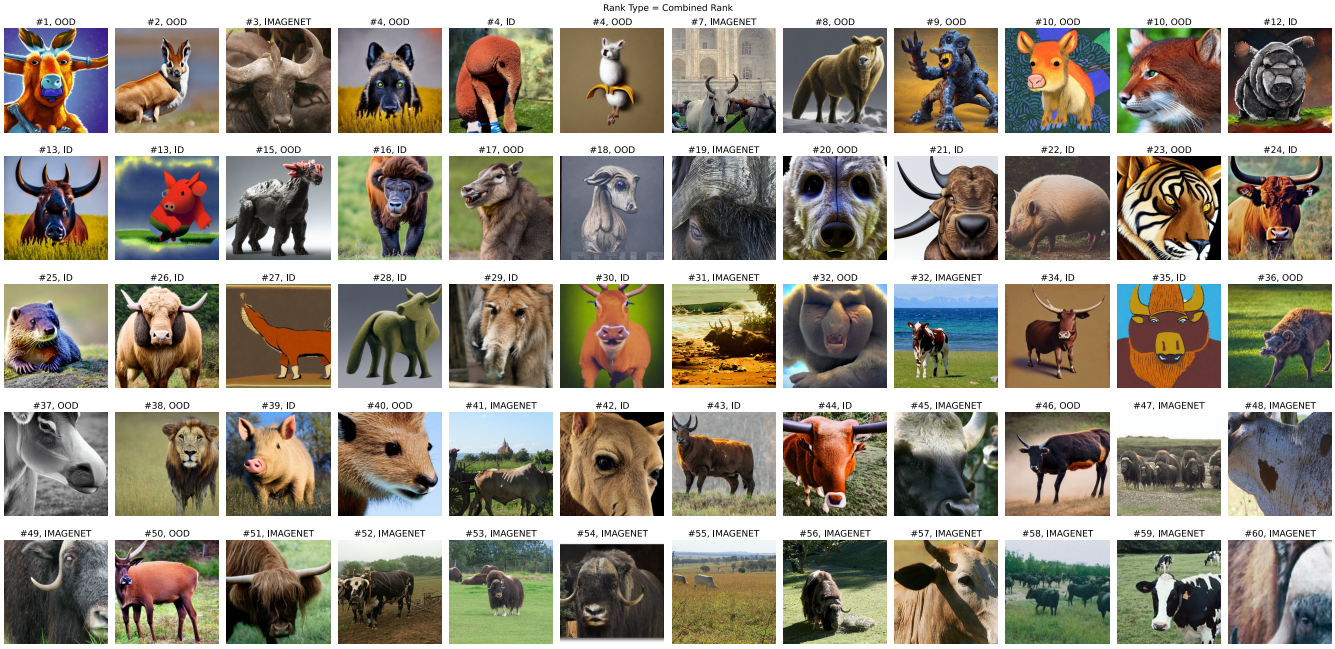


Figure 11: ELO Rankings Visualised (Combined Rank) (Zoom to read types)

6. Elo rankings

   - The box plots showed `OOD` & `ID` images to have lower (better) rankings than `IMAGENET` images.
   - For the `OOD` & `ID` images, in general the higher ranked show some aesthetically pleasing elements that do not directly relate to real world objects, they look nice by themselves, surprising and novel. The lower ranked images try to imitate something but do not succeed.
   - For the `IMAGENET` images the higher ranked images show some aesthetic value in colour, composition and focus. The lower ranked images have no focus on elements and look blurry.

7 Discussion and Future Work

The following five topics will discuss the five subquestions of the main research question by discussing interpretation of the results and looking at the limitations of the results gathered.

7.1 Type of creativity

In the methodology section (Section 4.1), we already concluded that `Dream-OOD` could potentially show exploratory creativity, and that at this point it would not be able to show transformative creativity, mainly because it lacks the ability to evaluate its own results and change its output accordingly.

7.2 Ranking creativity

With the user study and the Elo score calculation, we have shown that it is possible to rank images in a consistent manner using the three creativity features, the significance of this was shown by table 9.

It has to be noted that participants have probably found it difficult to distinguish the Novel and Surprising features in images, since there is a lot of agreement in those categories in all image types as is demonstrated by Figure 7 and 6.

There are differences in the ranking, which depend on which creativity features are used, so for future research it would be a good idea to keep on testing whether all three creativity features give important information or rather biased information. Although Boden's framework works really well

in describing creativity, distinguishing those features in a specific subset of creativity (images) might be too difficult for participants.

In the user study, participants were shown random pairs of images, instead of basing the pairings on current Elo scores. In chess, match-ups of players are directed by the Elo scores of the players. Using the logistic expected outcome formula (Section 2.3), the overall Elo scores in chess follow a logistic distribution. As shown in Figure 8, the Elo scores calculated from the creativity votes still follow this logistic distribution.

However, a problem that can arise is that of selective pairing. In chess this problem happens arises when players purposefully choose to only match up with what they think are overrated players, and avoid to play against underrated players. Even though the Elo scores of players should not over/underrated, it is still possible merely because of cold-start problems with new players, lucky match-ups, or just lucky games. Although the images in the user-study could not choose with which images they were paired, there could still be some luck involved in the pairing, especially because of the cold start of the Elo scores, all starting at 1500.

7.3 A model that appears creative

We can says with certainty that the images of the `OOD` & `ID` type appear more novel, surprising and valuable than the images from the `IMAGENET` type. What that in turn says about the creativity weighs heavily on the assumption of Boden's creativity framework. Although it is not possible to draw exact conclusions from the framework, the assumption in this study was that the three features, Novelty, Surprise, and Value, could describe creativity.

According to this assumption, the ranked images show that the images from the `OOD` & `ID` type which were generated by the `DREAM-OOD` model appear more creative than the images of the `IMAGENET` type.

Some limitations of this conclusion are that a single image class was chosen, the `Ox` class. So in reality we can only make this conclusion for this class, further research has to be done to use multiple different classes, also using a larger image sample size.

Lastly, the `IMAGENET` database is not an art database. The goal of this research was to go over Boden's Lovelace questions, which answered whether it was possible for a programme to appear creative, in general. It was not the goal to test whether a programme could be more creative than human creativity. This, however, is also an interesting topic to research.

7.4 Features of creativity

We have found some features that looked like they could have an effect ont the final ranking of the images, however, no stastical measures have been used to go over these features. Further research is required to find more features that influence the way people perceive creativity in images.

7.5 Being vs appearing creative

Given that Dream-OOD operates within the structured conceptual space of the text-conditioned latent space of a diffusion model, its creativity aligns with Boden's definition of exploratory creativity. Although the generated images are novel and surprising, they adhere to the underlying structure and constraints of the model.

However, whether Dream-OOD is truly creative or merely appears to be, is a philosophical question. Although the model can generate novel and surprising images, it lacks intentionally and does not possess the self-awareness to evaluate its creations. Therefore, while Dream-OOD demonstrates exploratory creativity within its defined space, it does not exhibit the full range of creative capabilities associated with human artists.

8 Responsible Research

To ensure that this research is done ethically and with responsibility, multiple precautions were taken to make sure it remained ethical.

8.1 Human Research Ethics

The user-study went through a thorough procedure of Human Research Ethics. This process involved a detailed assessment of potential risks and ethical considerations associated with the research, ensuring the protection of participants' rights and well-being.

Key ethical measures implemented in the study include:

- **Informed Consent:** Participants where shown an opening statement (Appendix A) detailing the study's purpose, procedures, and their right to withdraw at any time.
- **Anonymity:** The study was designed to be completely anonymous and without the collection of personal data.
- **Age Restriction:** An age criterion of 18+ was implemented to prevent the participation of minors, in accordance with ethical guidelines for research involving human subjects.
- **Transparency:** The opening statement informed the participants about the use of AI-generated images and non-generated images.
- **Data Handling:** The collection of votes happened on a secure TU Delft owned server, mitigating tracking of participants. The data itself could only be accessed by the researchers.

8.2 Calculation & Discussion of results

The research findings, including the collected votes, Elo rankings, evaluated results, and the Jupyter notebook that processed the results. will be published in/together with a paper that will be made publicly available in the TU Delft research repository.

References

[1] M. Boden, "Creativity and artificial intelligence," vol. 103, no. 1, pp. 347–356.

[2] X. Du, Y. Sun, X. Zhu, and Y. Li, "Dream the impossible: Outlier imagination with diffusion models," vol. 36. ISSN: 1049-5258.

[3] M. Boden, *The creative mind: Myths and mechanisms: Second edition*. Pages: 344.

[4] M. Boden, "Creativity: A framework for research," vol. 17, no. 3, pp. 558–570.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. ISSN: 1063-6919.

[6] A. E. Elo, *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation. Google-Books-ID: onUazQEACAAJ.

[7] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, "Is temperature the creativity parameter of large language models?."

A Opening Statement for participants

The opening statement that a participant would see before voting on pairs of images. It was also available in Dutch.

A.1 Participating in research

Dear participant,

You are being invited to participate in a research study titled Contemporary Creativity: The Many Faces of AI Art. This study is being carried out by Jari de Keijzer, a Bachelor student in Computer Science and Engineering at Delft University of Technology, under supervision of Dr. Anna Lukina.

The purpose of this research is to discover whether images that are generated by a specific Artificial Intelligence (AI) model can appear to be creative. To answer this question, you, as participant, will vote on the creativity of pairs of pictures. For each pair of images, you will vote on the image that you find the most novel, surprising or valuable. This means that for each pair of images you will enter three votes. The pairs of images will contain both AI generated as non-generated images. The AI images will have both more realistic (inlier) and stranger (outlier) images. During the voting process you will NOT be shown what category the image belongs too, this is part of the research. Participating in this research will take you about 15 minutes to complete.

This research is completely anonymous, so NO personal data will be collected. This voting website is hosted on a TU Delft owned server, on which no other data than your anonymous votes will be collected. These votes will be used to rank the images according to an ELO score (explained on the next page). With this ranking the creativity of the AI model will be evaluated. The collected votes, the ELO ranking, and the evaluated results will be published and presented in a paper which will be publicly available on the TU Delft research repository.

Anyone of the age of 18 years or older is allowed to participate in this research. It is important for this research to get a diverse collection of votes. More participants will increase the accuracy of the ranking. On that note, you, the participant, are allowed to share this research with others.

Your participation in this study is entirely voluntary and you can withdraw at any time. Since the votes are completely anonymous, they cannot be linked to you, which makes deleting 'your' votes impossible after submission.

Continuing to the next screen will start the voting process, which means you will be an anonymous participant.

Thank you for your time and participation,
Jari de Keijzer

B Explanation for participants

The explanation boxes that a participant would see before voting on pairs of images. It was also available in Dutch.

B.1 Vote by using your intuition!

Measuring creativity in photos - How does it work?

1. You will see two photos
2. Make three choices per pair of photos:
   - Which photo is the most novel?
   - Which photo is the most surprising?
   - Which photo is the most valuable?
3. Click on the buttons between the photos
4. Choose based on impulse and intuition
5. Click the green 'Submit' button
6. The next two photos will appear
7. You are asked to vote 30 times
8. Participating takes about 15 minutes

B.2 What is novelty?

An image is more novel than the other if it contains more new elements that you have not seen before. These new elements do not have to be surprising or valuable.

B.3 What is surprising?

An image is more surprising than the other if it has an unexpected twist with, for example, strange combinations of different elements. This surprising element does not have to be new or valuable.

B.4 What is valuable?

An image is more valuable than the other if it has aesthetic (what it looks like) or emotional value. This can be, for example, a beautiful composition or an image that moves you. This value does not have to be new or surprising.

B.5 What is an ELO rating?

An ELO-rating (score) is a numerical representation of a player's strength. It is most commonly used in chess and checkers but can be applied to any game where players compete one-on-one. In this research, the images are the 'players', instead of people. You, as a participant, determine who the winner of the 'game' is. After enough votes, a ranking of photos is created, with the most creative image having the highest score.