Faculty of Science
Human-Computer Interaction

Bachelor Thesis

Gaze-driven discrimination of Computer Graphics from Photo Images

Eberhard Karls Universität Tübingen
Faculty of Science
Wilhelm-Schickard-Institute for Computer Science
Human-Computer Interaction
Clara Riedmiller, clara.riedmiller@student.uni-tuebingen.de, 2022

Time period:	22.03.2022 - 02.08.2022
Supervisor:	Prof. Dr. Enkelejda Kasneci, University of Tuebingen
Advisor:	Dr. Efe Bozkir, University of Tuebingen

Abstract

The computer generated faces we encounter nowadays are on a constant increase regarding amount and quality. It is therefore important to investigate how humans perceive and interact with them. We exposed subjects to photographs of real faces and computer generated images with different levels of realness. They viewed the images first and, in a second phase, classified them based on whether they believed them to be of real or synthetic origin. We used an eye tracker to monitor subjects' eye movements in order to make inferences about their cognitive load while completing the tasks. We furthermore investigated what underlying factors in subjects' personal backgrounds potentially influenced their behavior.

Contents

1 Introduction

Over the last few years, the amount of fake content on the internet has been steadily increasing [1]. Especially manipulating videos and images has become more accurate and faster with improving technology and an increasing amount of data to sample from. Deep fakes superimpose a face of a person onto a video of another, thereby creating content of the person ”saying and doing things they did not say or do” [7].

This ability humans have created can be misused in malicious ways, thereby posing a ”serious threat to knowledge” [7]. Just recently, a deep fake of Ukraine’s president surrendering to Russia during the 2022 Russia-Ukraine war has gone viral online [28]. This highlights the importance of regaining control over the abilities this technology has given to a wide array of people. One way to tackle this is to educate people on what distinguishable features to look for when detecting computer generated content, which Institutions such as the German government aim at with postings on their websites [3]. Another way of regulating fake content is to build better classifiers; i.e. technology that detects and flags this information on the internet. Some projects with this goal are also listed in the German Government’s site, which they are investing into [3].

In order to contribute to both these approaches, it might be crucial to investigate how humans interact with computer generated content. Understanding how well humans are able to discriminate between real and computer generated content, as well as trying to understand the underlying cognitive processes, could help in building better detection algorithms, that flag and label false information as such.

To understand how this knowledge can be applied, we need to look further into the how generative adversarial networks (GANs) function. A GAN is a machine learning algorithm, that uses deep learning in order to generate new instances of a specific object. The concept was first introduced by Ian Goodfellow et al. in 2014 [10] and has since found a diverse range of applications. Output of GANs can range from text, to speech, to images (photographs to art), basically anything that exists in the real world. This allows for creation of fake content that can be distributed via the internet, by training a GAN to write fake news articles or produce images and videos depicting something that did not happen in the real world. A GAN’s architecture consists of two neural networks: a generator and a discriminator. The generator creates new instances of an object by looking at a training dataset and learning the defining features of the object's instances. When looking at images of faces, in a simplified manner this would mean: Looking at a large amount of photographs of real faces, learning that each has eyes, a nose, ears, hair, etc. ..., then learning what each of the features approximately look like and applying this conceptual knowledge to create new instances of these photographs. This content gets then passed on to the discriminator network, which decides if it is accurate enough to pass as an image of the original training dataset. In our case, the discriminator decides if this newly-generated image of a face looks like it could be a photograph of a real face. These networks compete against each other until reaching sufficiently good results and the discriminator is no longer able to distinguish the images from the training data set and images created by the generator. Images that pass this test then form the GAN's output.

Now, the question arises: How do humans benefit from increasingly more believable computer generated content, when even today's quality poses problems? The short answer is: Building better GANs can help detect false information. Paper [46] explored detecting fake news stories using a network that was trained to create them. Their GAN titled "Grover" can produce news stories humans perceive as more trustworthy than fake news written by another human. When tasking Grover with the detection of fake news, it was able to do so with 92% accuracy for stories produced by itself. Even fake news written by humans could be detected with 73% accuracy this way. Several other approaches to tackling this issue with in similar ways are listed in article [43].

Investigating human behavior while they are detecting computer generated content could help improve a GAN's discriminator network. By better understanding human visual attention during this task, discriminator networks could be alerted to certain features of an image or text that humans can still detect as being markers of its synthetic origin. A recent example of this general approach can be seen in paper [18], which factors in how users interact with real and fake content in order to improve fake news detection.

Though so far, we have only mentioned the negative ramifications of this technology, creating realistic digital humans can also have a lot of positive applications. During the Covid-Pandemic, several aspects of our lives have found their way into the virtual realm. Similar to playing a character in a video game, embodying digital avatars can help us have a more immersive experience as actors in virtual meetings, classrooms (see [8]) and social settings.

When it comes to designing these avatars, we should take into consideration the cognitive impacts observing human-like objects with varying degrees of realness has on the user. Upon first consideration, one might think: "the more human, the better!". The uncanny valley theory (see [22]) suggests otherwise. While there is a monotonous increase of affinity from 0 up to a local maximum of 70%, a human's affinity related to human-likeness towards that object drops down to negative levels after, suggesting an aversion (see fig. 1). The global maximum is at 100% human likeness, represented by a healthy person. This could suggest that when creating digital avatars or computer-generated faces, we should either aim at obviously fake or convincingly real ones in order to avoid giving the viewer an unsettling feeling or increase their cognitive load by evoking a cognitive conflict [44].

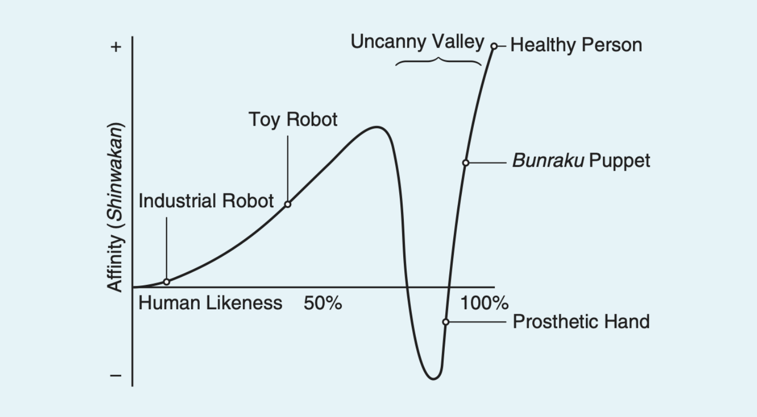


Figure 1. The uncanny valley. A human's affinity towards an object in relation to its human likeliness[1].

Several different approaches to this issue can be observed in the industry: Facebook's metaverse (see [20]) allows humans to interact online by taking the form of obviously fake, comic-book-like looking characters. Similarly, Microsoft Mesh (an addition that will be made to Microsoft Teams, a virtual meeting platform, see [21]) will offer avatars that mimic our facial gestures and move according to the users speech. Epic Games on the other hand has created MetaHuman (see [6]), an avatar-creation software that uses their own "Unreal Engine" and aims at creating photo-realistic avatars in real-time. This technology will be accessible to everyone and can be used in varying projects.

[1] © [2012] IEEE. Reprinted, with permission, from authors Masahiro Mori, Karl F. MacDorman and Norri Kageki. "The Uncanny Valley [From the Field]", IEEE Robotics & Automation Magazine, 06/2012

1.1 Scope of this thesis

As motivated above, it is increasingly relevant to investigate how humans interact with computer generated faces. In this study, we want to gain insight into a human's cognitive processes when observing these images and detecting their synthetic origin.

We therefore measured our subjects' responses to computer generated images of faces with different levels of realness. In the process, we investigated how well they are able to discriminate between real and fake content, as well as recording their visual behavior simultaneously in order to gain insight into the underlying cognitive processes. As a non-invasive way of monitoring cognitive processes, we used an eye tracker to record specific eye movement measurements. Analyzing these can provide an indication of subjects' cognitive loads [45].

To sum up the cognitive load theory (see [31]) briefly: The working memory has limited capacity. Tasks that demand the same skill and therefore same area of the working memory cannot be solved at the same time. The cognitive load relates to how much of this limited capacity is already occupied. If there is a high cognitive load, few resources are left for processing additional tasks.

In addition to looking at the subject's cognitive load, we inquired other self-perceived measures about the subject's personal background using questionnaires. This investigates if there are underlying factors that alter the subject's behavior.

Subjects were further split into two groups. When viewing the images, one group was primed on the origin of the image (computer generated/ real), while the other was not made aware of this. This indicates whether labeling content that is computer generated has effects in cognitive load (1). After the viewing, subjects were objected to the same images, this time discriminating the images on the basis of whether they seemed real and computer generated. In this second phase, we can observe whether the first group of subjects can recall the image condition when encountering the image a second time (2), as well as getting and unbiased assessment by the second group (3).

Results that can be inferred from this data can be split into three categories:

1. **Rating**: How well were subjects able to discriminate between real and computer generated images? What was their decision based on?

2. **Eye Movements**: Are there differences in eye movements (such as gaze patterns or pupil dilation) when viewing and rating computer generated or real images?

3. **Interrelations**: How do the above mentioned measurements relate to each other? Are there underlying personal factors (such as experience with deep fakes), that influence a subject's behavior?

1.2 Related Work

Previous research in the field of computer generated faces has investigated discrimination between computer generated and real images [17] and general eye tracking measurements such as gaze spread during this classification [4]. More in-depth eye movement analysis has been conducted for classifying fake and real news stories [47]. We aim to combine these approaches in our study.

The paper ”More Real than Real: A Study on Human Visual Perception of Synthetic Faces” (see [17]) produced a data set containing computer generated and real images of human faces. The computer generated images stemmed from three different GANs, each producing images with increasing levels of realness. Upon viewing, subjects rated each image on a 7-point scale, thereby deciding if they believed the image was real or computer generated. By analyzing subjects' responses, the paper investigated how the realness of the image impacted their accuracy and confidence.

An in-depth recording and analysis of eye movements was conducted in the paper ”Fake-NewsPerception: An eye movement dataset on the perceived believability of news stories” (see [47]). The focus here was to investigate how people recognize fake news stories. Relating eye tracking to perceived believability of the news articles can give an insight into ongoing cognitive processes while analyzing the content.

Classification of computer generated images of faces and eye tracking were combined in paper [4] with the title ”Using Eye-tracking to Study the Authenticity of Images Produced by Generative Adversarial Networks”. In addition to having a similar experiment setup to paper [17], an eye tracker was used to record subjects' visual behavior during the rating. The analysis of this data was focused on patterns in subjects' gaze.

The "Cognitive Conflict as Possible Origin of the Uncanny Valley" was explored by paper [44]. Faces with different levels of human likeness (these were created by step-wise morphing of a robot image to a human face) were presented to the subjects, which then had to classify the as human or non-human as fast as possible. By analyzing movement of the computer mouse, authors concluded there is an additional cognitive cost to processing images that are in the region of around 70% human likeness. They supposed this was due "cognitive conflict as evoked by simultaneous activation of two categories".

While eye tracking and the above mentioned mouse tracking are a non-invasive and convenient way of monitoring cognitive processes during image classification, brain-imaging techniques can provide further insight. Paper [23] found that while subjects' verbal feedback indicated they could not differentiate between real and computer generated images, their brain signals indicated otherwise. In the latter, there were significant differences in activity when observing real vs computer generated images. Our study will relate subject's rating behavior to their cognitive load and investigate whether the effects can be replicated using an eye tracker.

Our experiment setup is most similar to the paper [4]. In addition to recording discrimination behavior and tracking subjects' eye movements while completing this task, we added a viewing phase, where we recorded subjects' eye movement's when solely observing the image. Subjects were further split into two groups; one was primed on the origin of the image before viewing it ("computer generated"/ "real"), while one was left naive to this fact and the topic of the experiment. Classifying the images was handled in the second phase, where we recorded subjects' rating answers as well as their eye movements. When analyzing eye tracking data, we focused on its implication on the cognitive load subjects experienced during the tasks. We further did not limit the time subjects could take for each task (in contrast to [17] & [4]) and measured viewing and rating time as dependent variables. We implemented the option of providing a written answer in the rating where subjects could mention distinct features of each image that indicated to them whether it was real or fake.

We conducted this exploratory pilot study with the goal expand the findings already established by the above mentioned papers. Effects and possible tendencies in our data that could motivate future research.

2 Method

2.1 Subjects

We recruited 22 subjects in total, who were split up into two groups. Both consisted of 7 men and 4 women ($64\%m, 36\%f$), their mean ages being 27.82 ($SD = 9.84$) for group 1 and 26.36 ($SD = 9.17$) for group 2. Subjects were recruited from personal and work contacts and received no compensation. They were informed about data collection, usage and storage before starting the experiment. All subjects were naive to the purpose of the experiment.

Concerning their eyesight, subjects were instructed before the experiment to wear glasses or contact lenses if needed. All of them reported having normal or corrected vision during the Eyesight Questionnaire (section 2.4.2).

As subjects were recruited from personal and work contacts, we wanted to note and report how much their occupation was related to computer science. Subjects rated how much their current or past occupations are related to Computer Science on a 5-point scale ranging from ”Not at all” to ”A little” to ”A lot” (1-5 points). Overall, subjects reported a mean value of 3.55 ($SD = 1.65$) with no significant difference between the group 1 ($M = 4.0, SD = 1.61$) and group 2 ($M = 3.09, SD = 1.64$), ($t(20) = 1.311, p = .205$).

2.2 Groups

Subjects were split into two groups. Group 1 was primed on the image condition, while group 2 saw the images unbiased. Group 2 was also left naive to the topic of the experiment until after they had completed the viewing phase (see fig. 4). To their knowledge, they were simply looking at images of faces.

2.3 Apparatus

Gaze data was recorded using a Tobii Pro Spectrum Eye Tracker [41]. By detecting eyes with two cameras, it is able to track the subject's pupils with high accuracy despite small head movements. This eliminated the need for a chin rest and thereby allows for a more natural setting [41].

The eye tracker was fixed below and calibrated to a 1920x1080 screen, which subjects were placed in front of. After assuming a comfortable seating position, they were told to move as little as possible during the Viewing Phase and the Rating Phase fig. 4, where their gaze data was recorded. This ensured a proper distance from subject to screen and enabled steady eye tracking.

During the whole experiment, lighting conditions were controlled by blocking light from the windows and closing the curtains. We made this decision based on our test runs, where the calibration provided much better results in very low-light conditions (not no-light, there was still some controlled illumination coming from the screen). Tobii's light precision quality report did not show any reduction in data quality for low-light-conditions [36].

Before each phase in which gaze data was recorded, we performed a calibration. The 5-point calibration we used is part of Tobii's Eye-Tracker Manager. In addition to adjusting the eye tracker's data to changes in individual pupil geometry, it also includes a visual indication as to whether the subjects distance to the screen was appropriate. The optimal range for this distance lies between 55 and 75cm [29].

Recording the gaze data was managed by our application (see section 2.8). Data was sampled at a frequency of 600Hz. We recorded left and right eye coordinates relative to the screen, left and right pupil dilation and the validity of these measurements respectively. Every sample was equipped with a system timestamp, as well as a time of day timestamp.

2.4 Questionnaires

2.4.1 Demographic Questionnaire

Subjects stated their personal information in a demographic questionnaire. This included their age, their gender (m/f/d) and whether they had any kind of refractive error (near-/ farsightedness). If they answered yes on the last question, the Eyesight Questionnaire (section 2.4.2) would follow after.

Subjects also reported their basic political stance on 10-point scale. The scale ranged from "left" over "center" to "right". This measurement was recorded analogously to the fake news paper [47].

2.4.2 Eyesight Questionnaire

The eyesight questionnaire inquired more information about the subject's vision in case they had a refractive error. Though the Tobii Pro Spectrum eye tracker we used does not have problems detecting pupils through conventional glasses or contact lenses, there are some factors that can potentially negatively influence the quality of the eye-tracking data ([37], [38]). Subjects answered the following questions:

- Are you wearing anything to correct your eyesight?
  - if glasses: Are your glasses NIR blocking? (NIR = near-infrared light)
  - if contact lenses: Are your contact lenses bifocal or varifocal? (Not the same prescription throughout)
- Is your prescription more than +/-6 ?(For at least one eye)

We mainly kept record of the data in case we ran into any otherwise inexplicable issues with data quality during the analysis (which luckily did not occur).

2.4.3 Familiarity Questionnaire

In the familiarity questionnaire, subjects answered questions about other factors related to computer generated images. More experience and familiarity with the topic could potentially alter subjects' rating and viewing behavior. The questions we asked were:

1. On average, how many hours a day do you spend on a computer?
   5-point scale (<1, 1-2, 2-4, 4-8, >8)

2. On average, how many hours a day do you spend on the internet?
   5-point scale (<1, 1-2, 2-4, 4-8, >8)

3. Are you familiar with deep fakes?
   5-point scale ("Not at all", "A little", "A lot")

4. Are your current or past occupations related to Computer Science?
   5-point scale ("Not at all", "A little", "A lot")

5. Please rate your own ability to recognize faces.
   5-point scale ("Worse than average", "Average", "Better than average")

Question (1) was posed analogously to paper [4], questions (3) and (5) were also posed in [17]. As a lot of people use their computer for work, without actually increasing the likelihood of being exposed to computer generated images, we asked subjects about their daily internet use separately (2). Lastly, question 4 was added to report potential bias of our subjects which could be introduces by recruiting subjects from personal and work contacts. We suspected there might be a higher link to computer science in our subject group than we would get when sampling the subjects from a population at random.

2.4.4 NASA TLX

After the Viewing Phase and the Rating Phase respectively, the subjects completed the NASA Task Load Index [11]. In order to provide an indication on the subjects' cognitive load while completing the task in each phase, this index records personal assessment of perceived stress-level in 6 dimensions.

How successful were you in accomplishing what you were asked to do?	Performance
How hard did you have to work to accomplish your level of performance?	Effort
How physically demanding was the task?	Physical Demand
How hurried or rushed was the pace of the task?	Temporal Demand
How mentally demanding was the task?	Mental Demand
How insecure, discouraged, irritated, stressed and annoyed were you?	Frustration

Table 1. NASA TLX. The six stress dimensions are listed on the right side, while the question inferring each is listed on the left.

Each question is answered on a 21-point scale ranging from "Very Low" to "Very High".

This initial questions of the NASA TLX are followed by a weighting procedure. Here, stress-dimensions undergo a pairwise comparison, where the subject decides which one of them they generally perceived as more significant. In order to remind subjects which stress dimension related to which question they had previously answered, we provided them a table (similar to table 1) they could refer to. Each value from the first part is then weighted by multiplying it by the number of times it was rated as more significant compared to another dimensions. All weighted values are added up, then normalized by dividing it by the total weighting. This final score gives an indication of the subject's overall cognitive load while completing the previous task.

2.5 Selection of Stimuli

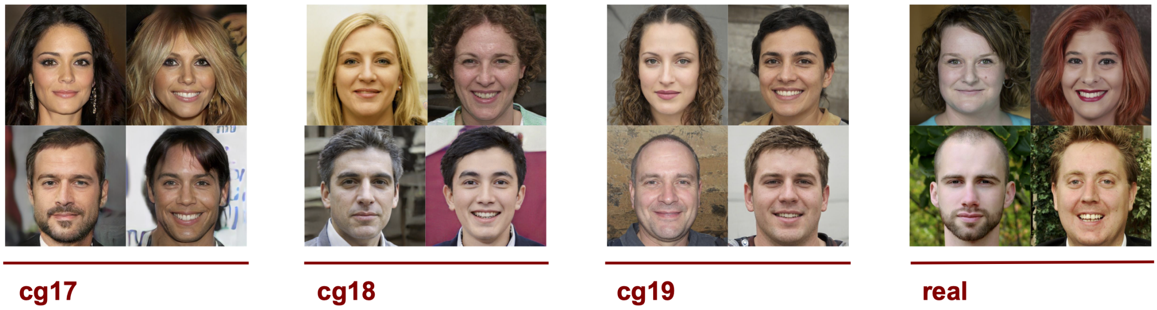


Figure 2. Exemplary subsample of images from each subcategory. Each image represents a combination of the three conditions c1, c2 and c3. From the top left to bottom right of each square: (female, closed mouth), (female, open mouth), (male, closed mouth), (male, open mouth)[2].

All computer generated and real images were produced by the authors of the paper ”More than Real: A Study on Human Visual Perception of Synthetic Faces” [17]. The data set contained 300 images in total. 150 of which real, 150 images were computer generated. The computer generated images were created by three different Generative Adversarial Networks (GANs). 50 each were generated by PGGAN from 2017 [13], StyleGAN from 2018 [14] and StyleGAN2 from 2019 [15]. As successors of the previous networks, each produced images with an increasing level of realness. In the following, I will be referring to origin of the images as image condition $c1$ with the factor levels ”cg17”, ”cg18”, ”cg19” and ”real”.

Due to our small number of participants, we used a smaller number of total images. Thereby making sure every image from our selection was seen by every subject, received the same number of ratings and had enough eye tracking data. This enables analyzing every image by itself and pointing out images that stand out in their condition. We therefore selected a subsample of 72 images from the 300 total provided by the paper. We first discarded images that did not fulfill our criteria, then removed the excess images by random selection.

[2] © [2022] IEEE. Reprinted, with permission, from authors Federica Lago, Cecilia Pasquini, Rainer Bohme, Helene Dumont, Valerie Goffaux and Giulia Boato. ”More Real Than Real: A Study on Human Visual Perception of Synthetic Faces”, IEEE Signal Processing Magazine, 01/2022.

The criteria we applied was based on unwanted patterns we discovered in the original dataset. First, we wanted to focus on the classification of faces and not the background. We eliminated images with backgrounds that would make it obvious whether they were real or computer generated.

Why is it bad using the background as the decisive feature? In general, looking at the background of images displaying a certain object often works for classification. An example of this is a simple classifier, that was tasked with discriminating between images of a husky or a wolf [26]. It reached satisfactory results in classifying the images but upon further investigation it appeared the decision was based on how much snow there was in the background of the image. This was not its intended purpose of classifying the animal. Similarly, in our experiment, we wanted subjects to focus on the face of the person displayed, not distinct pointers in the background, that would make it obvious the image was computer generated or real. This would influence not only the gaze patterns of the subject but also alter the information about humans' visual attention. As this research was motivated by being able to better detect deep fakes (which have a real background and only altered faces) and digital avatars (which can move around varying backgrounds), investigating how humans process the background of the images was redundant.

When deciding on which backgrounds we deemed permissible, we applied different criteria to the real and computer generated images, thereby eliminating patterns in the background that only appeared in one category or the other. In the husky-wolf example, this would be equivalent to removing images of huskies and wolves that had snow or a forest in the background, respectively. We would be left with animals that have a neutral background, such as a shared environment of both. Sorting the computer generated images, we looked for obvious processing mistakes in the background, such as geometric structures (see fig. 3, right). In real images, we sorted out ones with a very uniform background (only one color, no shadows), as well as images with clearly-defined patterns (see fig. 3, left).

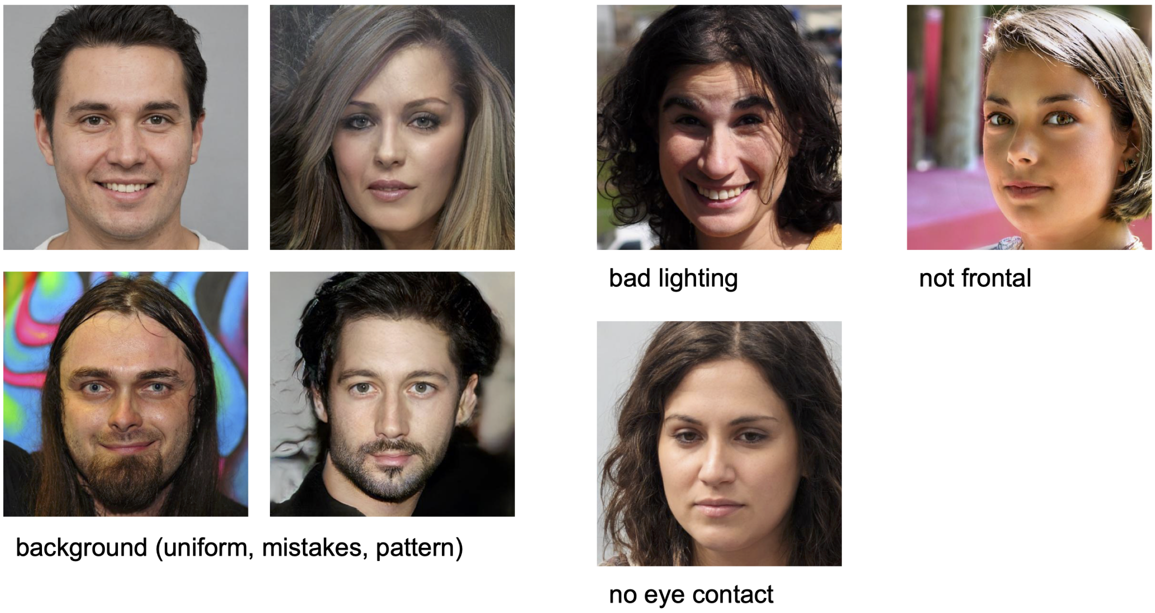


Figure 3. Images we eliminated from our dataset with reasons why.[3]

In addition to looking at the background, we also eliminated images with bad lighting/ harsh shadows (something that only appeared in real images), images that were not frontal (might influence symmetry perception) and faces that did not make eye contact with the viewer (might affect how much subjects look at different areas of the face). Examples all these other factors can be seen in fig. 3.

The 72 images we selected were chosen to be balanced for 3 conditions:

(c1) computer generated - real (cg17, cg18, cg19, real)

(c2) male - female (m, f)

(c3) open mouth - closed mouth (om, cm)

While paper [17] did record conditions $c1$ and $c2$, it did not factor in whether the image had a closed or open mouth ($c3$). We added this marker to the images manually based on how much teeth were showing. Specific facial features, such as teeth were often mentioned by subjects as a basis for their rating decision [17]. We recorded this to be able to analyze whether this had an effect on how much time subjects dedicated to looking at the mouth

[3] © [2022] IEEE. Reprinted, with permission, from authors Federica Lago, Cecilia Pasquini, Rainer Bohme, Helene Dumont, Valerie Goffaux and Giulia Boato. ”More Real Than Real: A Study on Human Visual Perception of Synthetic Faces”, IEEE Signal Processing Magazine, 01/2022.

and how much it influenced their rating decision. A lack or abundance of symmetry was also reported as a decisive feature, which is especially obvious in geometrical structural elements of the face like the teeth. We therefore wanted to record this additional image condition to see whether it had any effect on gaze patterns or subjects' ability to classify images correctly.

2.6 Sequence

Completing the experiment took around 45 minutes, though the duration varied between subjects, as the pace was self-regulated for all subtasks.

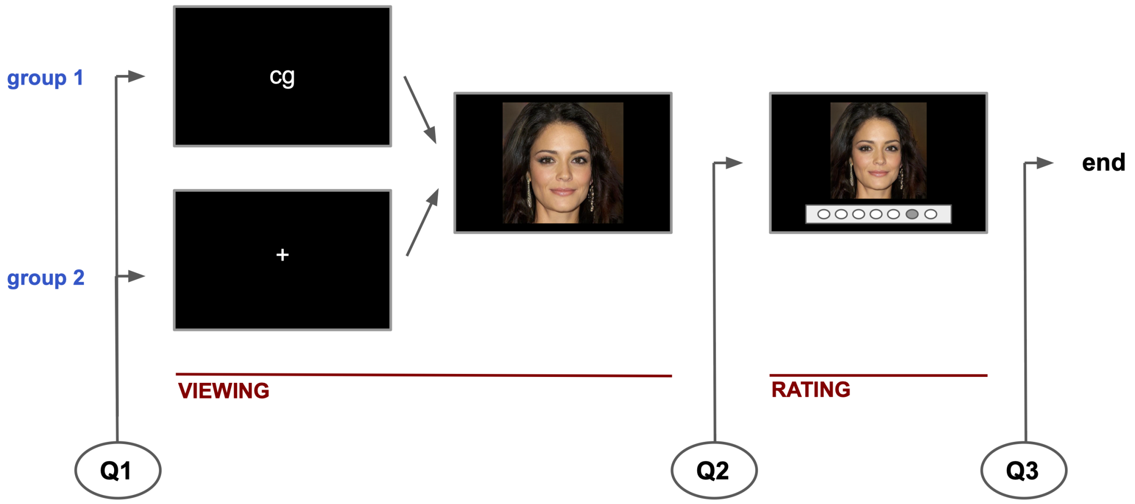


Figure 4. Sequence of the Experimental Design, from left to right. The experiment started off with **Q1**; the Demographic Questionnaire and if they reported a refractive Error, the Eyesight Questionnaire. After, subjects entered the Viewing Phase. They saw a sequence of images which was preceded by either a condition prime (real/ computer generated) for group 1 and a fixation cross for group 2. After the Viewing Phase, subjects completed the first NASA TLX and the Familiarity Questionnaire in **Q2**. Lastly, the Rating Phase was followed by **Q3**, the second NASA TLX, which concluded the experiment.

We split the experiment into two phases. Subjects cycled through all images twice, once only viewing them, then rating them in the second phase. Compared to rating images upon first seeing them, this had two advantages: Firstly, eye tracking data that was recorded as subjects observed the images was not influenced by them looking at the rating scale or typing text answers. Secondly, group 1 could be primed on the image conditions in the Viewing Phase without doing the rating at the same time. Group 2 could view the images unbiased in the first phase without knowing what they should be looking for (seeing the rating question would give them idea about the topic of the study). In order to see the initial reaction to the image in form of visual attention and behavior, gaze data would have to be recorded whilst only the image is being displayed, not the rating scale. Also, group 2 was left naive to the purpose of the experiment in order to get their unbiased reaction to seeing what to their knowledge was only faces. (not influenced by the rating question or the primes).

In **Q1**, subjects received instructions on how to navigate the user interface first, the completed the Demographic Questionnaire (section 2.4.1), optionally followed by the Eyesight Questionnaire (section 2.4.2), in case they reported a refractive error prior. After completing the calibration, they then received instructions on the task. Both groups had the task to observe the images and continue whenever they felt they had seen enough. While group 2 was not given further information, group 1 also was informed they would see before each image whether it was real or computer generated.

In the Viewing Phase, subjects saw the images in random order. Preceding the image was either a condition prime (group 1) or a fixation cross (group 2). Each was shown for 2 seconds (analogous to paper [4]).

The image was then displayed until the subject pressed the "Continue"-button or hit the "Return key" in the keyboard. This option was enabled after 2 seconds. We introduced the minimum viewing time in our experiment, as we suspected the task was posed in such a way the subject would have a hard time deciding how long is enough.

Images were shown 10cm by 10cm in the center of a black screen. On the bottom right corner of the screen, three arrows (">>>") indicated the "Continue"-button. After the 2 seconds minimum viewing time, it slightly changed hue from gray to white. This was only a subtle change to indicate to the subjects they now had the option to skip but at the same time not distract their visual attention.

After subjects had completed the Viewing Phase, they completed a second round of questionnaires **Q2**. The NASA TLX (section 2.4.4) to accessed their cognitive load while viewing the images. This was followed by the Familiarity Questionnaire (section 2.4.3), which was conducted after the Viewing in order to not give group 2 an indication about the topic of the study.

In the **Rating Phase**, subjects saw all images again while assessing the statement ”I think this image is computer generated”. On a 7-point scale, subjects rated from ”Completely Disagree” to ”Completely Agree” (Similar to paper [17]). They also had the option to elaborate on specific features of each image that helped them make their decision.

Additional instructions were given, as we wanted to avoid effects on gaze and cognitive load that were unrelated to the task at hand. For example, recalling information by trying to translate vocabulary to English from German or trying to remember what condition the images belonged to from the Rating Phase. We instructed subjects to write down the German word if they could not think of the English one. Group 1 was further instructed they could use information about the image condition in the rating if they remembered it from the viewing phase but to just proceed with the rating based on their own assessment otherwise. We did not limit the time subjects could take for the rating, as this might have discouraged them from providing text answers.

In **Q3**, subjects completed the NASA TLX a second time to assess their cognitive load for the rating phase.

Paper [4], that also investigated humans classifying between real and computer generated images, allowed subjects to end the rating session after whatever amount of images they wanted. In the end, the amount of images subjects saw ranges from 13 to 103 with a varying percentage of real to fake images. They investigated the training effect (participants getting better over time due to more exposure to the content) and the loss of accuracy over time (maybe due to fatigue). Weighing these against each other, they concluded findings ”may suggest that the ideal session duration or this type of experiment is between 20 and 40 trials” ([4], p.4). As in our experiment, we presented subjects with a fixed amount and selection of images (72 total, out of which 36 real and 36 computer generated), we applied these suggestions by splitting the trials into 3 blocks, containing 24 images each. Between the blocks, subjects had the option to take a break and continue whenever they felt ready to. In order not to influence participants by the order images were presented in, the blocks were balanced for condition c1 (24 total, 12 real, 4 cg17, 4 cg18, 4 cg19).

2.7 Data pre-processing

2.7.1 Perception Engineer's Toolkit

Raw eye tracking data was processed using the Perception Engineer's Toolkit [16]. In order to detect relevant events (fixations and saccades) in the raw eye-tracking data (see section section 2.3). Released in 2020 by Thomas Kuebler, this toolkit unites the different steps that need to be taken in processing Eye Tracking Data. Besides several different implementations of algorithms for event detection, it includes methods to reduce noise in the raw data as well.

In order to get our data into the correct format for the input, it was processed using pandas (v. 1.4.3). For each subject, two data frames were produced. One contained the eye tracking data from the viewing phase, one form the rating phase. As the Tobii Pro Spectrum Eye Tracker' gaze coordinates were reported relative to the screen, we converted them into pixels. We split the coordinate touples for the left and right eye's x and y coordinates into separate columns and added their validity.

We applied a pre-process gap fill to close small gaps in the data using linear interpolation. These small periods of missing data under 100ms are usually not due to blinking or looking away but rather temporary inference of eyelashes or other factors [40]. By setting the maximum gap length to be filled in this way to 22ms (default in both the Perception Engineers Toolkit [16] and Tobii Pro Lab [40]), we avoid losing information that might be relevant in the event detection following. For example, a fixation that contains a few missing samples could be split into two if this correction had not been applied before [40].

Raw gaze data also contains some noise [40]. This might occur due to the system setup or changes in the environment [40]. The median filter smoothes the data by iterating over each point and making its value the median of its neighbors in a time window of 30ms. Reducing the amount of noise is, for example, beneficial to fixation detection. In order to detect a fixation, using IVT (see below), coordinates of gaze data are being compared to each other. A fixation is detected whenever the coordinates are sufficiently close to each other.

From the several different implemented methods for event detection the toolkit offered, we chose Identification by Velocity Threshold (IVT). The original method from Salvucci Goldberg (2000) [27] was in this implementation extended by Tobii's Implementation of

IVT Fixation Filter [35]. IVT defines a velocity threshold in °/s. Point-to-point velocities are being computed, after which velocities above the defined threshold are classified as saccades, whilst the ones below are classified as fixations [2]. We used a minimum fixation duration of 60ms and a velocity threshold of 30°/s.

The output we received from this processing were separate files for each trial. Each file therefore contained one subject observing one image in one phase. For each sample, the toolkit added whether it was detected as being part of a fixation or a saccade and the index of this event. Also computed were fixation and saccade summaries for each trial. Per trial, we got the following measurements:

- average fixation duration
- standard deviation of the fixation duration
- number of fixations
- fixation rate
- within fixations standard deviation
- within fixations x-range
- within fixations y-range
- number of saccades
- average saccade duration
- saccade rate
- average saccade amplitude
- hv ratio
- average saccade velocity
- average peak saccade velocity
- average time to saccade peak

In order to save this data and automate the processing for every subject, we modified the source code of this toolkit. We also added functions to compile a list of all fixations that were made, which would be relevant to detecting areas of interest in further analysis.

2.7.2 Compiling the data sets

In order to conduct the analysis, we compiled a data set containing all eye tracking information, as well as image conditions and subject information.

We finished processing the eye tracking data by computing the average pupil dilation using pandas (v. 1.4.3) and numpy (v. 1.22.3) in python (v. 3.9.12). Also using pandas, we compiled all information into one big data frame. This included the gaze data from the viewing and rating phases, image-specific information (image number, conditions,...) and subject-specific information (subject number, group, gender, NASA TLX scores, demographic information, questionnaire answers,...).

Subject's rating answers were also further processed. For each subject, we computed an overall score, a score for each category c1, an overall score including confidence, a realness rate and pure confidence. The overall scores by giving them +1 point for a correct , 0 for a neutral and -1 for an incorrect answer. Scores per category were calculated the same way this time only adding up over images from cg17, cg18, cg19 and real respectively. The overall score with confidence took into account additionally, how close to either end of the scale the answer was. The realness rate reported what percentage of images were rated as real by the subject. Pure confidence was computed by taking the absolute value of the rating answer with confidence without taking into account the correctness (how far on either end of the scale is the answer).

In the end, we produced the above mentioned data set (1), where each cell represented one trial. This was then processed further for our analysis by calculating mean values for the measurements per subject, per image condition c1, per phase (2). This data set of means was the one used in our analysis. The final data set we produced was a list of all fixations (3) made by each subject, for each image in both phases. This reported the fixations' start and end time, their duration, centroid x and y coordinate, standard deviation and x and y range.

2.7.3 Processing text answers

We applied word correction to text answers (Example: "backround" $->$ "background") using textblob (v 0.16.0), a python package for processing textual data [32]. We also removed fill words and prepositions ("the", "her", "his", "it",...), as well as punctuation marks (",", ".", "(", "("), manually for the word frequency histogram.

2.8 Application

2.8.1 Benefits of programming own application

A wide range of software available on the market today enables easy implementation of experiments conducted on a computer. To name a few examples; PsychoPy [25] is a cross-platform, open source package for implementing experiments. Tobii Pro Lab [34], is a Desktop Application that offers options to display an experiment on screen while automatically streaming data from the eye tracker, which can also be pre-processed and analyzed in the app. Both of these options though did not suffice to offer all features we required for our implementation, thus justifying the need to program an application from scratch.

Our requirements:

- implement and automate all application funcitonality (from questionnaires to viewing and rating phase), thus
  - eliminating the need for intervention by the experiment conductor. Thereby distracting subjects as little as possible.
  - not having to switch between screen and paper when answering questionnaires throughout the experiment sequence, in order not to strain the eyes or change visual behavior by switching between lighting conditions in the room.
  - control the amount of light coming off the screen especially during tracking sessions (minimizing reflections in eyes and glasses).
- stream gaze data from the eye tracker, organize data in a self-specified way. Recording which answers were given along with timestamps, as well as viewing and rating time.
- creating an intuitive and self-explanatory user interface and allowing interaction via radio buttons, text fields, mouse clicks or the keyboard.
- modular design:

- adapt the experiment sequence while developing the app, altering parts when necessary after the testing phase.
- enabling for future use of the basic framework of this application. Modules can be easily adapted or switched out.

Note: Continuing from a section to the other was possible via clicking the mouse or pressing the return key.

2.8.2 Build

The application was written in python (v. 3.6.13). The UI was implemented using tkinter (v. 8.6.11). The structure of the application puts great emphasis on modularity in order to adapt to a changing experimental design. The modules were all contained in separate files, a simplified version of the application's logic and functionality can be seen in fig. 5.

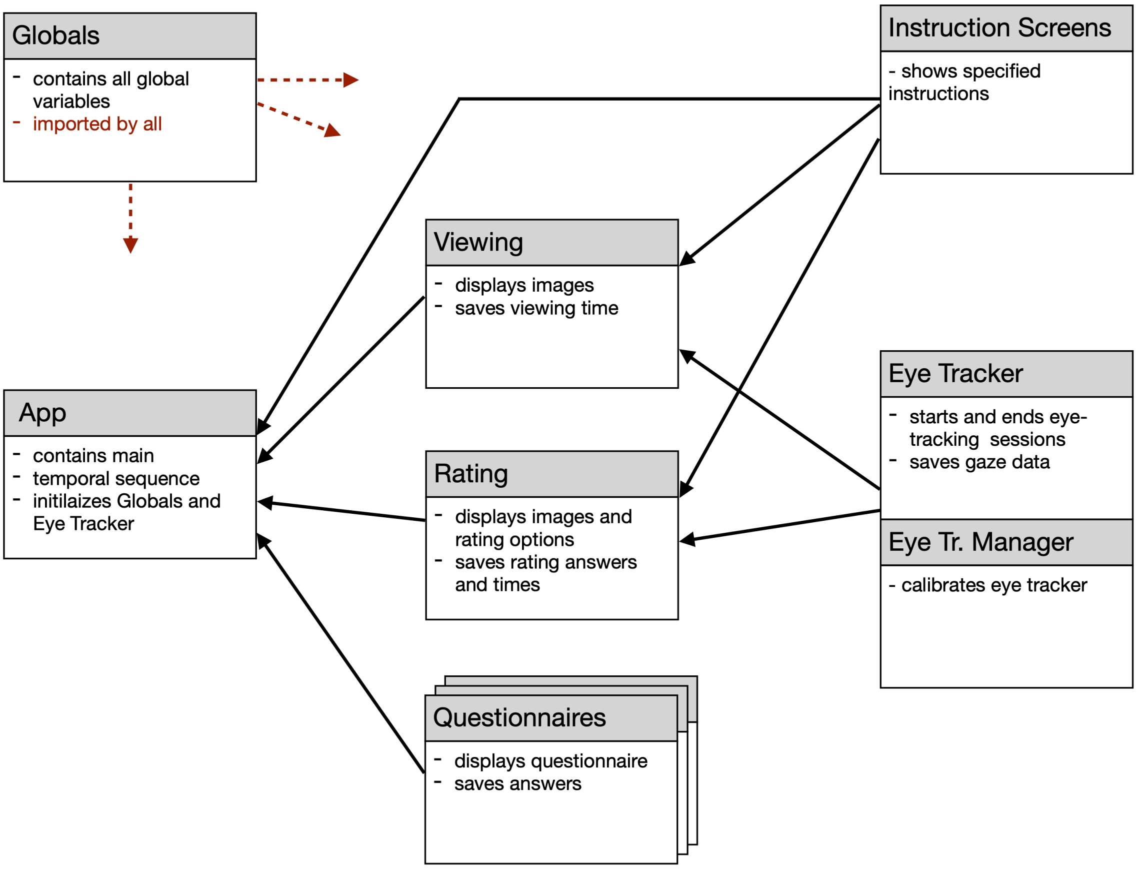


Figure 5. Simplified structure of the application. Boxes represent files (modules) with their key functions. Arrows represent one module being imported into and used by another module.

The temporal sequence was handled by the main **App** file. All other modules were imported into this one and called in order. This main sequence monitored global variables that changed from 0 (in progress) to 1 (done) after completing the section. The root window then displayed the next screen. For example: In the start of the experiment, an Instruction Screen containing a welcome message was displayed. After the subject had read the text and pressed "Continue", a global variable relating to the Instruction Screens incremented, therefore prompting the display of the Demographic Questionnaire.

Managing the temporal sequence very precisely was important to this experiment. Knowing exactly when a display was being called was relevant especially in parts of the experiment that were time-sensitive. Displaying an image before the user was actively able to observe it could have lead to mismatched timestamps and would have therefore resulted in faulty reaction times and gaze data.

In advance to starting the displays, the main App first retrieved the subject group and subject number from the command line input prompted by the experiment conductor. It then set all file paths in Global according to the subject number. A list of all stimulus images was compiled from the image folder and shuffled randomly for each person. This determined the order in which the subject would see the images in the Viewing and the Rating Phase. As both phases were each split into 3 blocks, we shuffled the images in a way that balanced each third of them for condition c1. By creating an instance of the Eye Tracker right in the beginning, we ensured the Eye Tracker was found and there was a stable connection. When this was not the case, the experiment conductor could interfere before the experiment had started, therefore not having to interrupt while data recording was in progress.

The **Globals** file contained all the global variables that were shared among the files. All variables could be accessed from every module. This was most convenient for the current subject number, the subject group and the subject-specific file paths, as each module had its own data saver. This is not best practice and should be implemented differently in the next version of the application. By having once central Data-Saver module that contains all file paths for example, the need for them to be accessible form everywhere could be eliminated. In our case though, this setup was convenient, as throughout the programming process, we added and modified a lot of variables. Without saving them globally, each function's input parameters would have to be altered with every change.

Images for the Viewing Phase were displayed in the **Viewing** module. After the minimum viewing time of 2000ms, the subject was able to continue to the next image by either clicking on the arrows in the lower right corner of a display or selecting this option using the "Return" key. Before displaying each image, the viewing module also produced the display for the inter stimulus interval (ISI) for 2000ms (analogously to paper [4]). Group 1 saw a condition prime before each image, while group 2 saw a fixation cross before each image. The eye tracking was active whenever there was an ISI or image showing and was paused whenever one block of images had been completed. It resumed when the next one started. Between the blocks, an instance of "Instruction Screen" was displayed, telling subjects to take a short break, which they could end at any point by clicking "Continue".

The **Rating** Module was structured analogously to the Viewing Module. Every subject saw the images in the same order as they did in the viewing phase, including the breaks. This time, they were instructed to rate them, as well as optionally providing a text answer. On a 7-point scale, subjects assessed the statement ”I think this image is computer generated.” from ”Completely Disagree” to ”Completely Agree”. Below the scale, there was a text field where subjects had the option to elaborate on specific features of the image that supported their decision. We prevented non-answers by implementing a pop-up window that reminded the subject to provide an answer before continuing and disabled the option to continue otherwise. Along with the self-determined rating time, the rating answer and the text answer were recorded and saved.

Instruction Screens displayed instructions consisting of a header, an instruction and the text that was displayed on the button. The button was placed on the lower right corner showed the next section in sequence upon activation. In most cases, it was labeled ”Continue” but also could give an indication to the next section, like ”Calibration” or ”Start”. The text for the header, the instruction and the button were stored in a text file. When another module required an instruction, it called Instruction Screen and passed the number of the instruction as a parameter. The number of the instruction corresponded to its line in the text file. We implemented it this way because the instructions contained the majority of the Application’s total text. If the experiment were to be conducted in another language in the future, changing the instruction text file would allow for an easy adaptation. This method of storing text could also be implemented for the questions displayed in the questionnaires and scale labels in the future.

Each of the **Questionnaires** (demographic questionnaire, eyesight questionnaire, familiarity questionnaire, NASA TLX and NASA TLX weighting) was assigned its own module. Keeping them separate allowed for more customization of each one. Depending on the questionnaire, answers were given by selecting a value in a scale with varying size or typing in a text answer field. Similarly to answers given in the rating, we avoided non-answers by only allowing subjects to continue only when they had answered all questions. Whenever this was not the case, the subject was reminded by a pop-up to check if they had given an answer to every question. Answers were saved to the specified location, the NASA TLX score was computed directly after the weighting.

Lastly, the **Eye Tracker** was based on Tobii’s guide for building python interfaces with their eye tracker [42]. In this module, we implemented methods to start and end the collection of gaze data by subscribing or unsubscribing to the data streamed from the eye tracker. These were called by the Viewing module and Rating module. Also in the Eye Tracker module, we adjusted the frequency of data sampling to 600Hz. From the gaze data provided by the eye tracker, we saved the gaze coordinates of the left and right eye, left and right pupil dilation, the validity of these four measurements respectively, and the system time stamp to separate files for Viewing and Rating.

The **Eye Tracker Manager** (ETM) was an adaption of Tobii's guide to integrating the Eye Tracker Manager into a python program [39]. It opened the ETM App at the desired points in the program (before starting the viewing and rating phase), where the experiment conductor supervised the calibration. After ensuring successful calibration, the "Continue" button prompted the next section to start.

3 Results

For the statistical analysis, we used t-tests and decided significance based on comparing the resulting p-values to an α-level of 0.05. We assumed normal distribution and homogeneity of variance. These assumptions will be especially valid, if we conduct the experiment with more subjects in the future, making the amount of subjects n=30 or larger. If these assumptions had not been given, we would have used the Mann-Whitney U test, which works for non-parametric distributions. As this, in some cases, produces smaller p-values, we decided against it. With this being a mainly exploratory pilot study, we wanted to find as many tendencies in the data as possible and not potentially miss out some interesting directions.

For each measurement reported in the results, we compared:

- differences between groups. Per image condition, per phase. Using independent t-tests.
- differences between image conditions within the same group. Per group, per phase. Using individual t-tests.[4]
- (if measurements were made separately on both phases) differences between phases. Per group, per image condition. Using paired t-tests.

As we made a total of 16 comparisons for measurement with one phase and 40 comparisons for measurements with two phases, we applied a multiple comparison correction. This prevents finding significant effects only due to the amounts of tests conducted. We used the Holm-Bonferroni method [12] for this, which compares more significant p-values to increasingly strict alpha-levels. Effects that were significant before the correction but not after are indicated in the plots as "*(ns)". We will still reported effects that become non-significant after the correction, as this exploratory study could give indications for follow-up research, which will benefit from seeing tendencies in the data.

Comparisons that did not produce significant p-values are not shown in the plots. Whenever there are no significance bars between the distributions, the above-mentioned

[4] We used individual t-tests here even though data stemmed from the same subjects. The initial data has unequal length, as there were 36 images from real but only 12 per computer generated image condition. This issue does not arise when working on the mean data set.

comparisons were still made for every measurement analysis that follows.

The plots were produced using seaborn (v. 0.11.2), an extension of the python package matblotlib and annotated using a self-modified version of the python package statannotations (v. 0.4.4). In the latter, we manipulated the test result outputs (automatically compute mean, standard deviation and report those along with the p and t value in APA standard). Furthermore, we adapted the source code to not display bars or significance indications for non-significant effects. As we conducted a large amount of comparisons, the plots would have been too crowded otherwise.

All plots that are shown in the following all follow the same systematic principles. When comparing distributions within one phase, the x-axis consists of 4 values, depicting the image conditions of c1 (cg17, cg18,cg19 and real). Furthermore, each condition shows two bars depicting distributions for each of the two subject groups. When a measurement was takes in the viewing and the rating phase individually, the graph doubles in width and repeats values on the x-axis a second time for the second phase. We hid outliers in order to get a more clean and comprehensive plots, especially after adding the significance bars.

3.1 Rating Answers

3.1.1 Scores

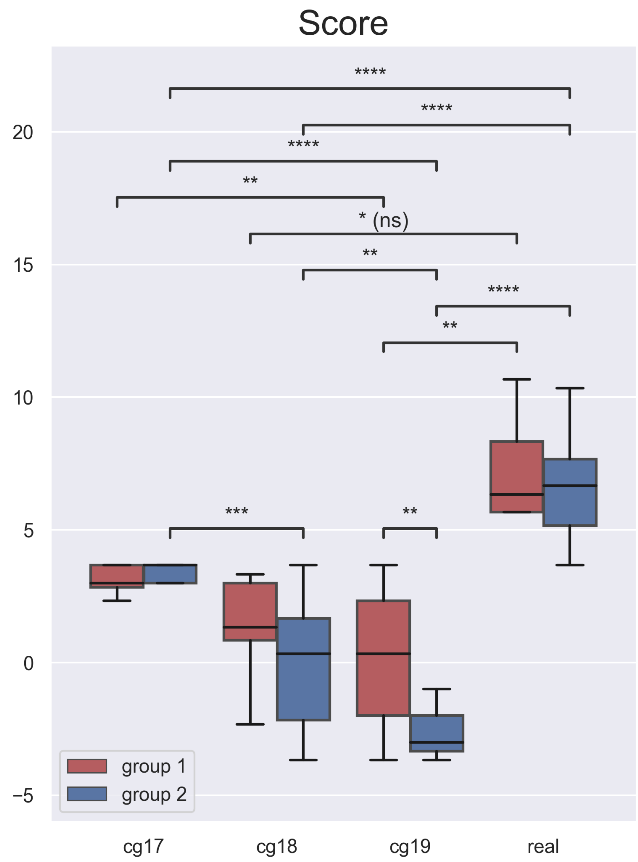


Figure 6. Average scores subjects achieved during the Rating Phase. Per group, per image condition. Stars indicate a significant p-value . (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

Subjects assessed the statement ”I think this image is computer generated” on a 7-point scale. This means answer values ranged from 1 to 7. An answer of 4, being the midpoint, was neutral, meaning subjects could not decide whether they agreed with the statement for this image. Values smaller than 4 meant the subject classified the images as real, while values greater than 4 meant the subject thought the image was computer generated.

Scores were calculated the following way: Subjects received +1 point for classifying the image correctly, -1 point for wrong classification and 0 for giving a neutral answer. These values were summed up for every image condition per subject, providing a mean score. The distribution of these can be seen in fig. 6. This computation did not take into consideration the confidence with which the answer was given (how close of either end of the scale was their answer placed, signifying their agreement with the statement), which will be discussed separately for the sake of better clarity.

Scores depicted this way can be interpreted as follows: Values that could be reached for each image condition ranged from +12 (subjects classified every image correctly), to a minimum of -12 (classified every image incorrectly). Scores greater than 0 indicate the subjects classified the images correctly more often than not, while scores below 0 indicate the opposite. Score that are approximately 0 demonstrate subjects classified only 50% of the images correctly, indicating the accuracy of their rating is equivalent to random choosing. This analysis resulted in the following findings:

Between the groups, though group 1 tends to be slightly better than group 2 for the computer generated images, there is only a significant difference in scores for images of condition "cg19". Here, subjects from group 1 ($M = 0.03, SD = 2.48$) performed significantly better in the rating than group 2 ($M = -2.58, SD = 0.98$), $t(20) = 3.244, p.004$.

Group 1 performed significantly worse for images from cg19 ($M = 0.03, SD = 2.48$) than for images from cg17 ($M = 2.79, SD = 1.44$), $t(20) = 3.191, p = 0.005$ and real images ($M = 5.58, SD = 4.83$), $t(20) = -3.390, p = 0.003$.
Group 2 showed significant differences in scores between all image conditions. Mean and standard deviation were distributes as follows: cg17 ($M = 3.27, SD = 0.61$), cg18 ($M = 0.0, SD = 2.48$), cg19 ($M = -2.58, SD = 0.98$) and real ($M = 6.64, SD = 1.93$). Subjects performed best when rating real images, scores in this category were significantly higher than (in descending order) for categories cg17 ($t(20) = -5.513, p = 2.139e - 04$), cg18 ($t(20) = -7.004, p > 0.0001$) and cg19 ($t(20) = -14.13, p < 0.0001$).

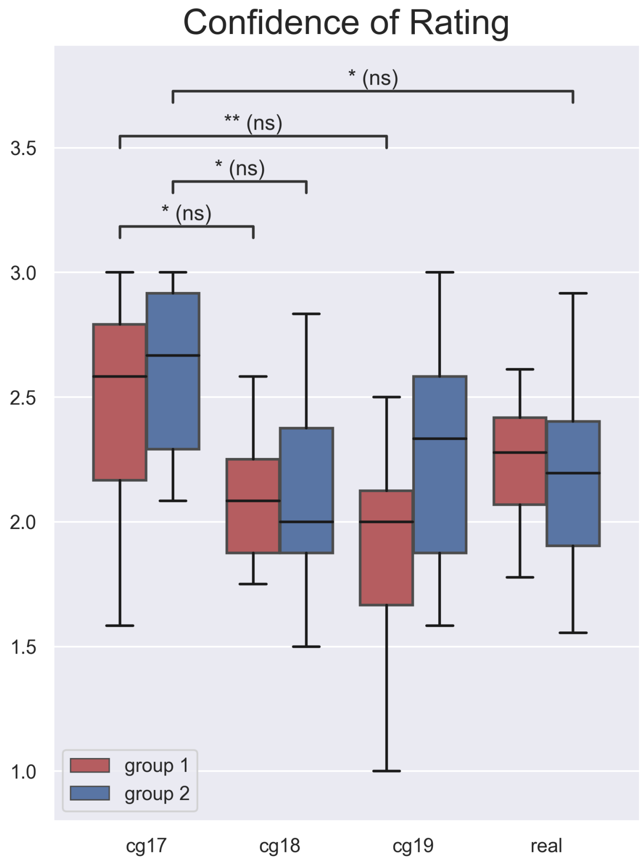


Figure 7. Average confidence subjects rated the images with, not taking correctness into account. Per group, per image condition. Stars indicate a significant p-value. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

When looking at only **Confidence** (fig. 7), effects were only significant before the multiple-comparison correction. Results were apparent as follows: For both group 1 and group 2, confidence was lower rating the images from $cg17$ than when rating images from $cg18$ (group1: ($cg17$: ($M = 2.45, SD = 0.47$), $cg18$: ($M = 1.99, SD = 0.46$), ($t(20) = 2.314, p = .031$)), group2: ($cg17$: ($M = 2.57, SD = 0.36$), $cg18$: ($M = 2.12, SD = 0.43$), ($t(20) = 2.663, p = .015$))). Group 1 also showed decreased confidence when rating images from $cg19$ compared to images from $cg17$ ($cg17$: ($M = 2.45, SD = 0.47$), $cg19$: ($M = 1.9, SD = 0.43$), ($t(20) = 2.855, p = .010$)). Group 2 showed decreased confidence when rating *real* images compared to images from $cg17$ ($cg17$: ($M = 2.57, SD = 0.36$), *real*: ($M = 2.16, SD = 0.41$), ($t(20) = 2.488, p = .022$)).

3.1.2 Rating time

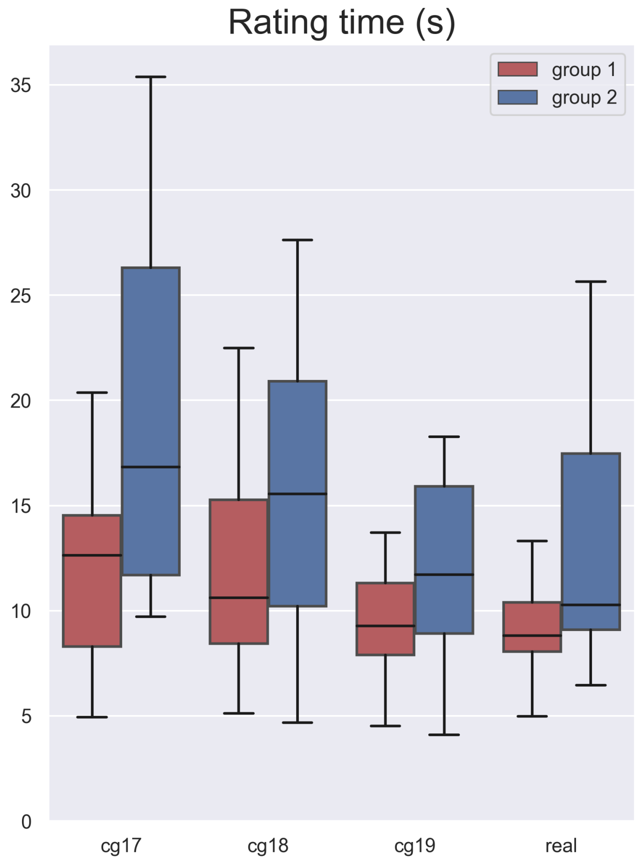


Figure 8. Average time subjects took for a rating. Per group, per image condition.

Comparing the **Rating time** (fig. 8), there were no significant effects when comparing the different distributions. It seems though there might be a tendency that subjects took the shortest amount of time when rating a real image with a linear increase the more computer generated the image was. Group 2 also had slightly higher mean values than group 1 for every category and seems to have a bigger variance. These effects might be better traceable in a study with more subjects and should be investigated further.

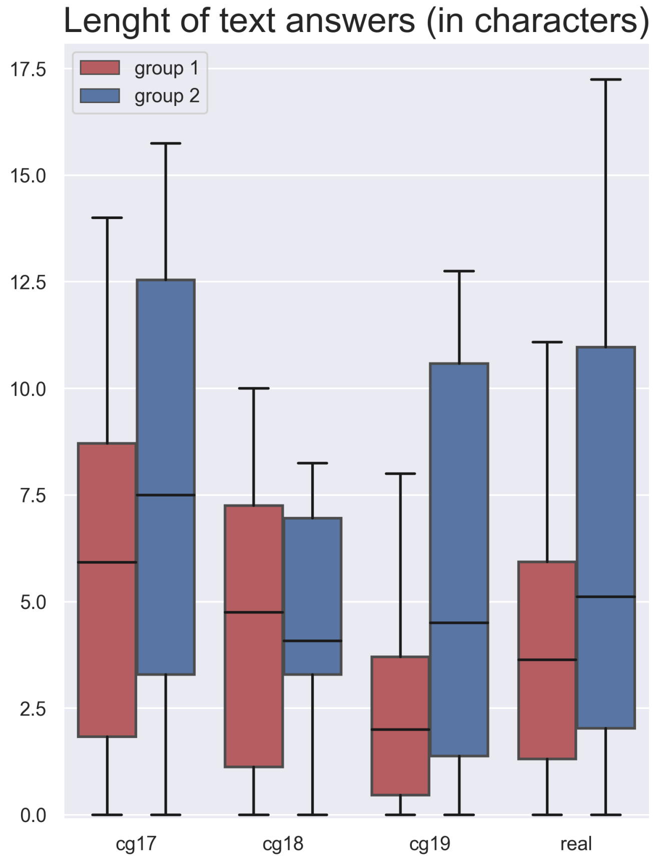


Figure 9. Average length of the text answer (in characters) given during the rating. Per group, per image condition.

Taking a longer time in the rating could not only be due to needing more time to process and categorize the image, but also relate to the amount of time it took to type out a text answer. Providing a written answer, where subjects could type out specific features of the image that supported their decision, was optional and we did not specify how long it could be. Therefore we also looked at the length of the text answer given, measured in number of characters, of the text answer. There were no significant differences in the **length of the text answer** (see fig. 9).

3.1.3 Viewing time

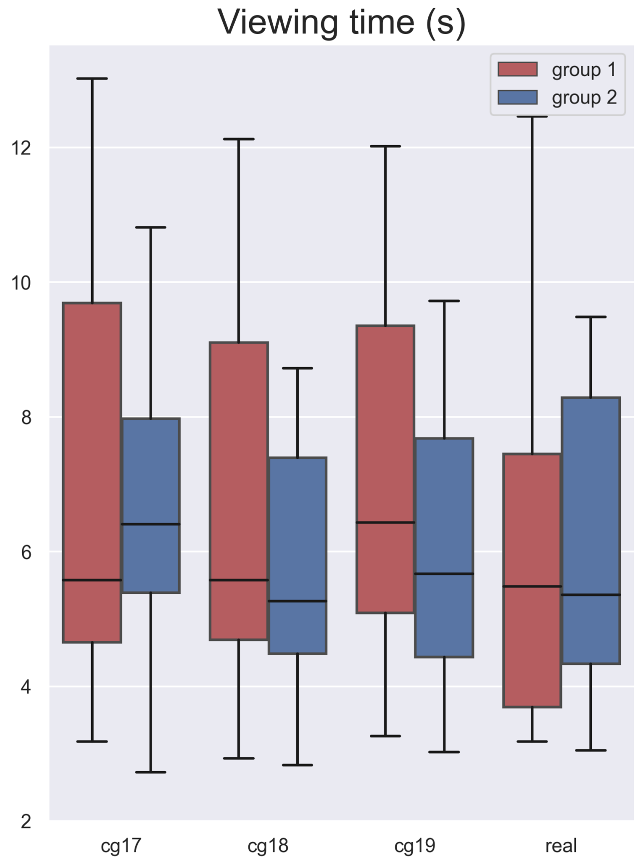


Figure 10. Average viewing time in seconds. Per group, per image condition.

The diagram in fig. 10 shows the average **Viewing time** per group. This indicates how long a subject viewed an image on average before continuing to the next one. There were no significant differences in group distributions.

To sum up the time and thoroughness subjects took for viewing and classification: We did not find any significant differences in distributions of rating time, viewing time, as well as number of fixations and saccades for both phases respectively. This indicates that even when giving subjects the freedom to take as much time as they need for either task, this seems to be independent of the image condition or their group.

3.1.4 Text Answers

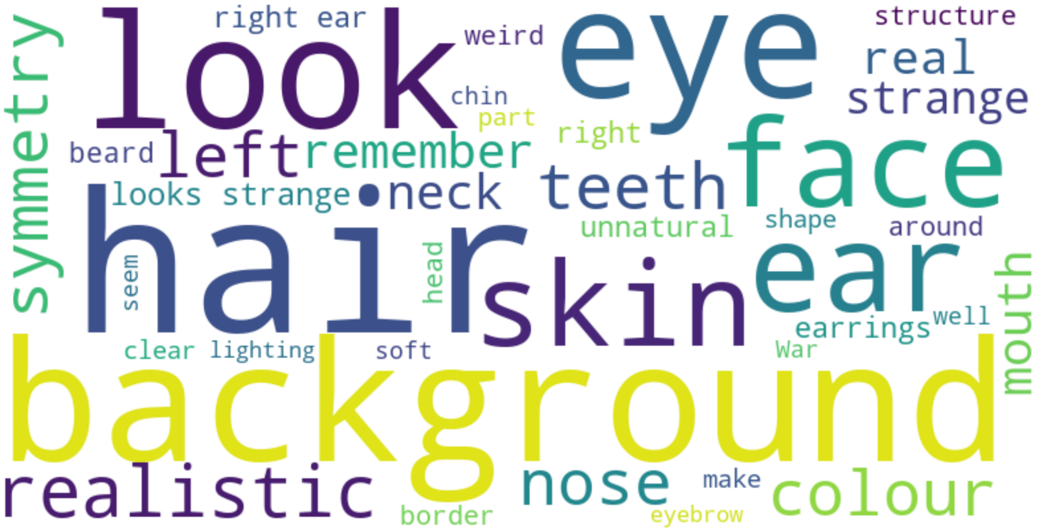


Figure 11. Word cloud showing the 40 most common text answers. Subjects had the option to describe specific features of each image that helped them make their rating decisions. Word size is proportional to its frequency.

In fig. 11, the **Word cloud** depicts the relative frequency of text answers. The 40 most common words were displayed in a font size proportional to how frequently they were mentioned by the subjects. This includes only text answers given in English; in order to include the German answers as well, one could translate the missing words. From a first glance at the data though, it seems only more complicated words were left in German, while subjects were able to produce more common words (like those seen in the cloud) in English.

This word cloud was created using wordcloud (v. 1.8.2.2) in python.

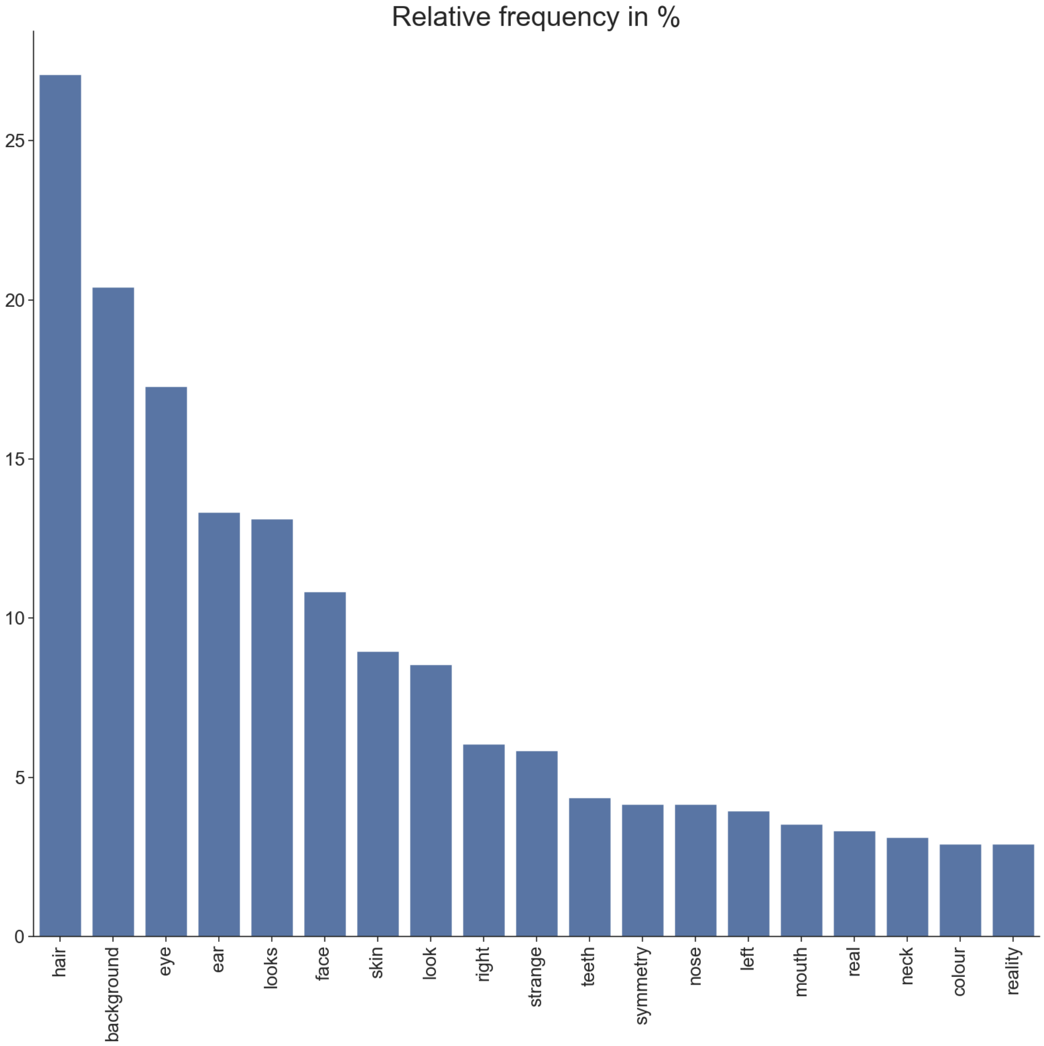


Figure 12. Relative frequency of the most common words. Depicted here is what percentage of the text answers given contained the word.

In total, subjects gave a text answer to 30.30% of all images (480 out of 1584). The histogram in fig. 12 shows how many of the text answers included the words on the x-axis. Hair was mentioned in 27.08% of text answers, background in 20.42%, eye in 17.29%, ear in 13.33%, looks in 13.12%, face in 10.83% and the remaining words in less than 10%.

3.2 Eye Movements

The motivation for collecting eye tracking data in this experiment is gaining insight into subjects' cognitive processes while conducting the viewing and rating of images. Analyzing eye movements can give an indication to the cognitive load a subject is experiencing while processing these tasks.

After compiling previous research in the field, the author of paper [45] proposes the following correlations:

The higher the cognitive load

- the longer the fixations
- the lower the fixation rate
- the longer the saccades
- the higher the saccade velocity
- the larger the pupil dilation
- the lower the blink rate
- the higher the blink latency

We will therefore look at these measurements in the following. Comparing differences in the distribution between groups, image conditions and phases can give an indication as to the underlying factors that prompt changes in cognitive load.

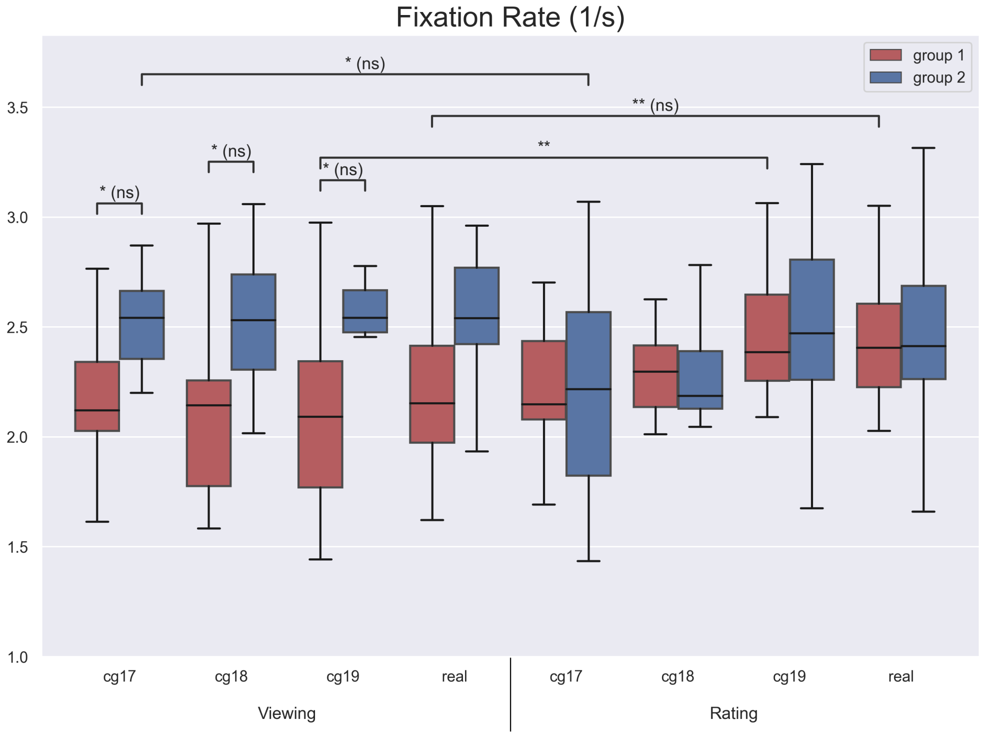


Figure 13. Average fixation rate (number of fixations per second). Per group, per image condition, per phase. Stars indicate a significant p-value. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

Differences in distributions of **Fixation Rate** fig. 13:

For all cg conditions, group 1 had a significantly lower fixation rate than group 2 ($cg17$: (group1: ($M = 2.18, SD = 0.46$), group2: ($M = 2.51, SD = 0.21$), ($t(20) = -2.152, p = .044$)), $cg18$: (group1: ($M = 2.11, SD = 0.44$), group2: ($M = 2.53, SD = 0.32$), ($t(20) = -2.535, p = .020$)), $cg19$: (group1: ($M = 2.11, SD = 0.48$), group2: ($M = 2.51, SD = 0.24$), ($t(20) = -2.491, p = .022$))).

Group 1 had a lower fixation rate for images from $cg19$ in the Viewing Phase ($M = 2.11, SD = 0.48$) than the Rating Phase ($M = 2.46, SD = 0.31$), ($t(20) = -4.507, p = .001$). The same was true for real images (V: ($M = 2.18, SD = 0.52$), R: ($M = 2.43, SD = 0.31$)), ($t(20) = -3.203, p = .009$).

Group 2 had a lower fixation rate in the rating phase compared to the viewing phase for images from $cg17$ (V: ($M = 2.51, SD = 0.21$), R: ($M = 2.19, SD = 0.53$)), ($t(20) = 2.421, p = .036$).

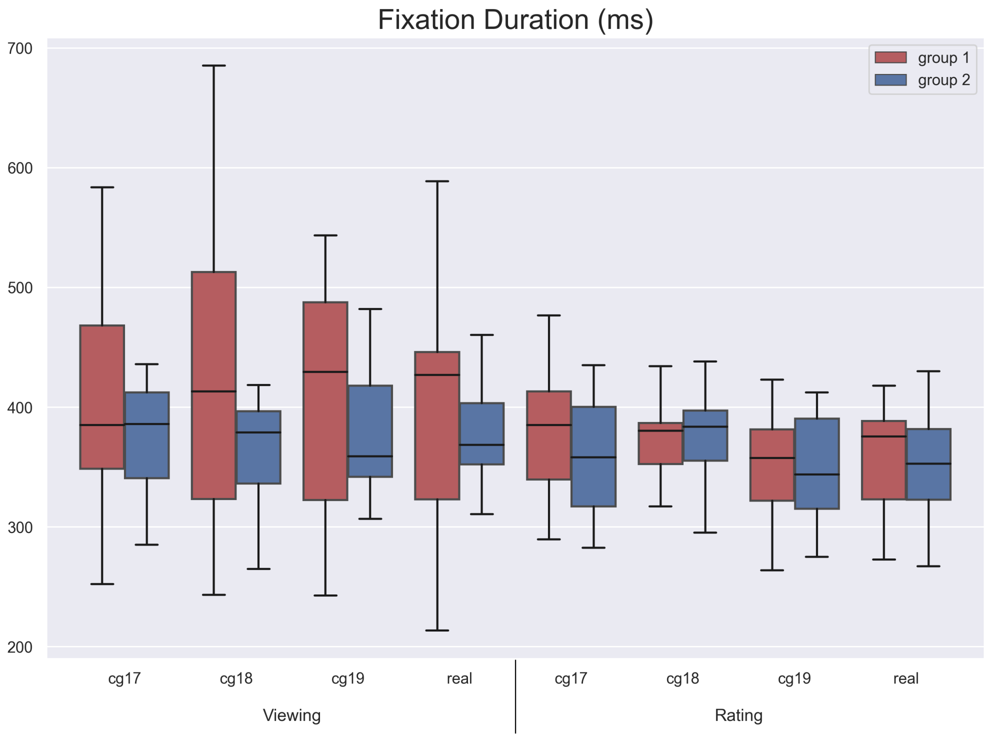


Figure 14. Average fixation duration in ms. Per group, per image condition, per phase.

There were no significant differences in **Fixation duration** (see fig. 15) when applying t-tests to compare the distributions' means. There seems to be a larger spread (larger variance) in the Viewing phase for group 1. This should be investigated using additional testing going forward.

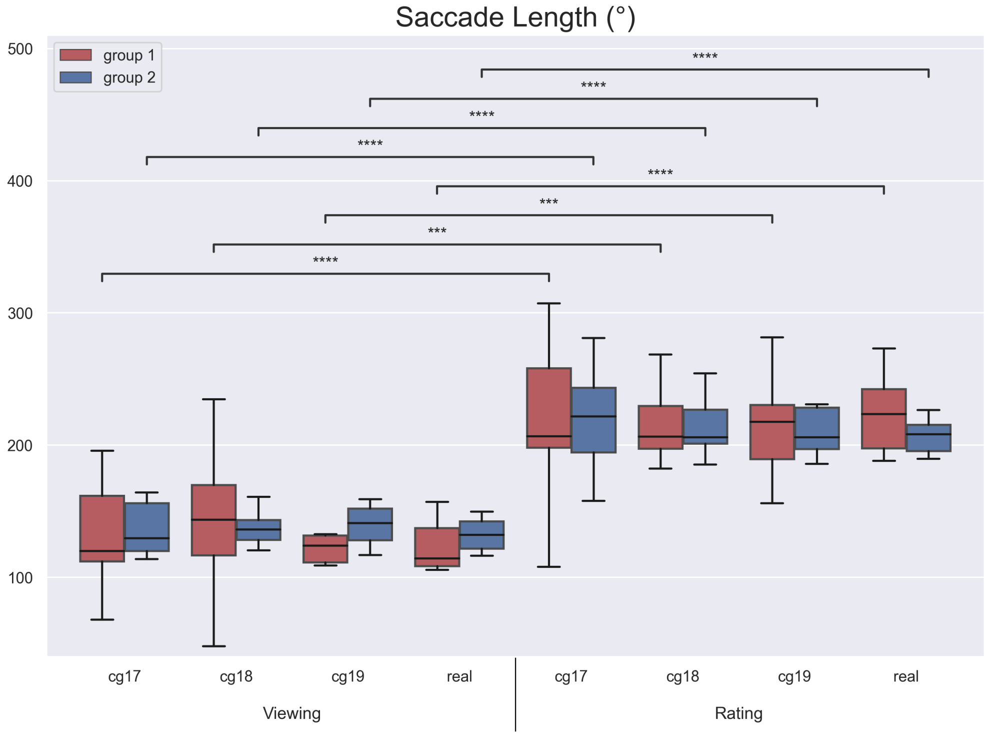


Figure 15. Average length of saccades in visual angle degree. Per group, per image condition, per phase. Stars indicate a significant p-value. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

Both groups had a larger **saccade length** in the rating phase compared to the viewing phase. This effect was visible for both groups and all conditions. Testing provided results as follows: Group 1: ($cg17$: (V: ($M = 132.08, SD = 37.83$), R: ($M = 223.19, SD = 55.87$), ($t(20) = -6.480, p < .0001$)), $cg18$: (V: ($M = 142.06, SD = 48.98$) , R: ($M = 210.56, SD = 55.26$), ($t(20) = -5.940, p < .001$)), $cg19$: (V: ($M = 128.78, SD = 43.34$) , R: ($M = 208.44, SD = 61.14$), ($t(20) = -3.406, p = .007$)), *real*: (V: ($M = 121.34, SD = 39.21$), R: ($M = 210.47, SD = 57.78$), ($t(20) = -9.858, p < .0001$))), Group 2: ($cg17$: (V: ($M = 136.5, SD = 19.89$), R: ($M = 219.06, SD = 35.53$), ($t(20) = -11.569, p < .0001$)), $cg18$: (V: ($M = 136.75, SD = 12.64$), R: ($M = 220.9, SD = 34.32$), ($t(20) = -8.021, p < .0001$)), $cg19$: (V: ($M = 139.48, SD = 15.45$), R: ($M = 219.8, SD = 37.0$), ($t(20) = -7.984, p < .0001$)), *real*: (V: ($M = 132.29, SD = 12.02$), R: ($M = 213.18, SD = 26.94$), ($t(20) = -9.462, p < .0001$))).

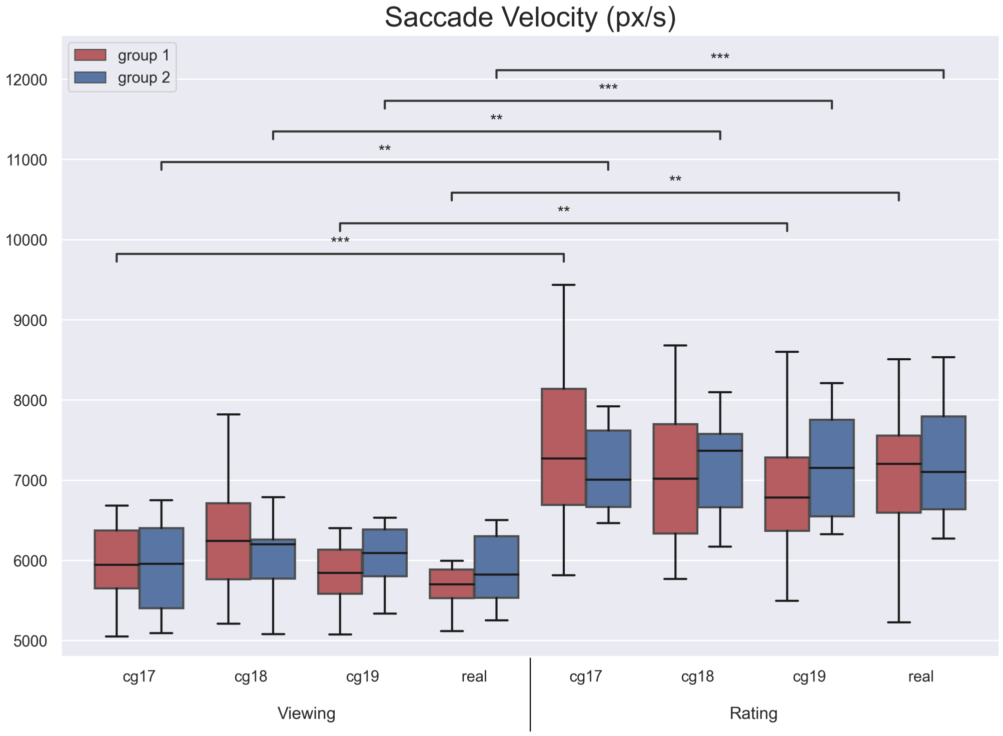


Figure 16. Average saccade velocity in pixels per second. Per group, per image condition, per phase. Stars indicate a significant p-value. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

The average **Saccade velocity** was significantly higher in the viewing phase than in the rating phase. This applied to both groups and all conditions, except for group1 $cg18$.

Comparisons provided results as follows: Group 1: ($cg17$: (V: ($M = 6037.4, SD = 700.78$), R: ($M = 7432.99, SD = 1151.12$), ($t(20) = -5.360, p < .001$)), $cg19$: (V: ($M = 5983.6, SD = 802.63$), R: ($M = 6934.35, SD = 1034.93$), ($t(20) = -3.406, p = .007$)), $real$: (V: ($M = 5965.1, SD = 895.76$), R: ($M = 7108.72, SD = 954.58$), ($t(20) = -3.222, p = .009$))), Group 2: ($cg17$: (V: ($M = 5899.45, SD = 598.45$), R: ($M = 7656.16, SD = 1684.0$), ($t(20) = -3.915, p = .003$)), $cg18$: (V: ($M = 6037.14, SD = 534.72$), R: ($M = 7466.67, SD = 1293.87$), ($t(20) = -4.223, p = .002$)), $cg19$: (V: ($M = 6075.32, SD = 382.68$), R: ($M = 7212.86, SD = 697.83$), ($t(20) = -5.607, p < .001$)), $real$: (V: ($M = 5904.12, SD = 451.44$), R: ($M = 7203.75, SD = 755.53$), ($t(20) = -5.062, p < .001$))).

Note: For analyzing the saccade velocity, we used the output of perception engineers toolkit (section 2.7.1). This lacked proper documentation on the unit of saccade velocity.

After some investigation, we assumed based on the scale of the vales that the unit must be pixels per second. Using a visual angle calculator [30], we computed the conversion of pixels to degree. With screen dimensions measurements of 24 inches [41] and a subject-screen distance of 60cm (optimal range: 55cm to 75cm [29]), we concluded to the following conversion rate: $38px \approx 1°$. Meaning dividing these values by 38 will give them in degrees per second ($6000\frac{px}{s} \approx 158°, 7000\frac{px}{s} \approx 184°$). This does not have an impact on the effects depicted here, as it is only a scaling factor.

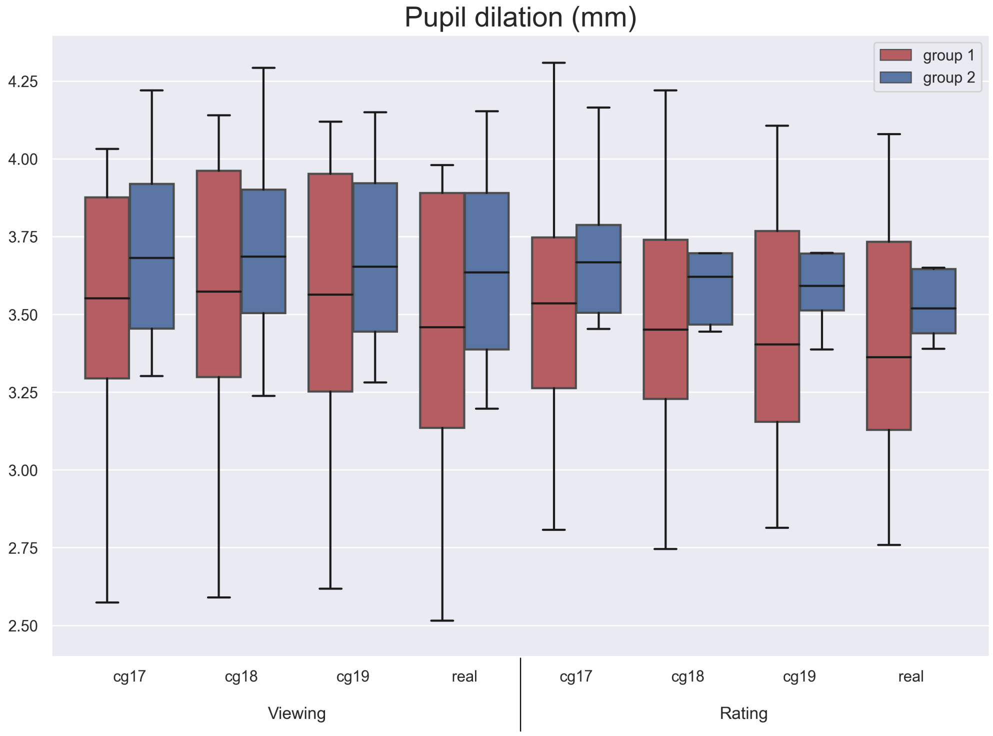


Figure 17. Average pupil dilation in mm. Per group, per image condition, per phase.

As another eye movement measurement indicating changes in cognitive load, subjects' average **Pupil dilation** is shown in fig. 17. No significant effects appear when comparing the distributions' means using t-tests, but there seems to be tendencies in the variance. In both the viewing and the rating phase, subjects from group 1 seem to have much larger variance than group 2. By applying a baseline correction, we could investigate whether this is due to individual differences in the subjects or due to changes in cognitive load.

3.3 Interrelation

This section is dedicated to investigating general relationships between the data.

Firstly, we investigate correlations between different factors. Besides looking at the interrelation of different eye movement measurements, we looked into whether underlying factors of a subject's personal background have an effect on their viewing and rating behavior.

Secondly, we analyze the distribution of rating answers and whether these have effects on eye movement measures or rating behavior. This allows looking at the data from another angle; not focusing on what images condition the images stemmed from but how real or fake the subjects believed it to be.

3.3.1 Correlations

We computed correlations between the measurements using pearson-correlation coefficients. These values should be taken with a grain of salt, as repeating this operation on large amounts of data can quickly produce a big number of significant correlations. For a proper analysis in the future, it would be reasonable to break the mean data set down more. We decided to still include and discuss the correlations we found in order to investigate tendencies in the data that could be interesting looking into in follow-up research.

The correlation coefficients here were computed on the basis of the mean data frame. Eye movement data was not separated by phases, meaning it includes data from both. In this calculation, p-values were smaller than 0.001 for all effects.

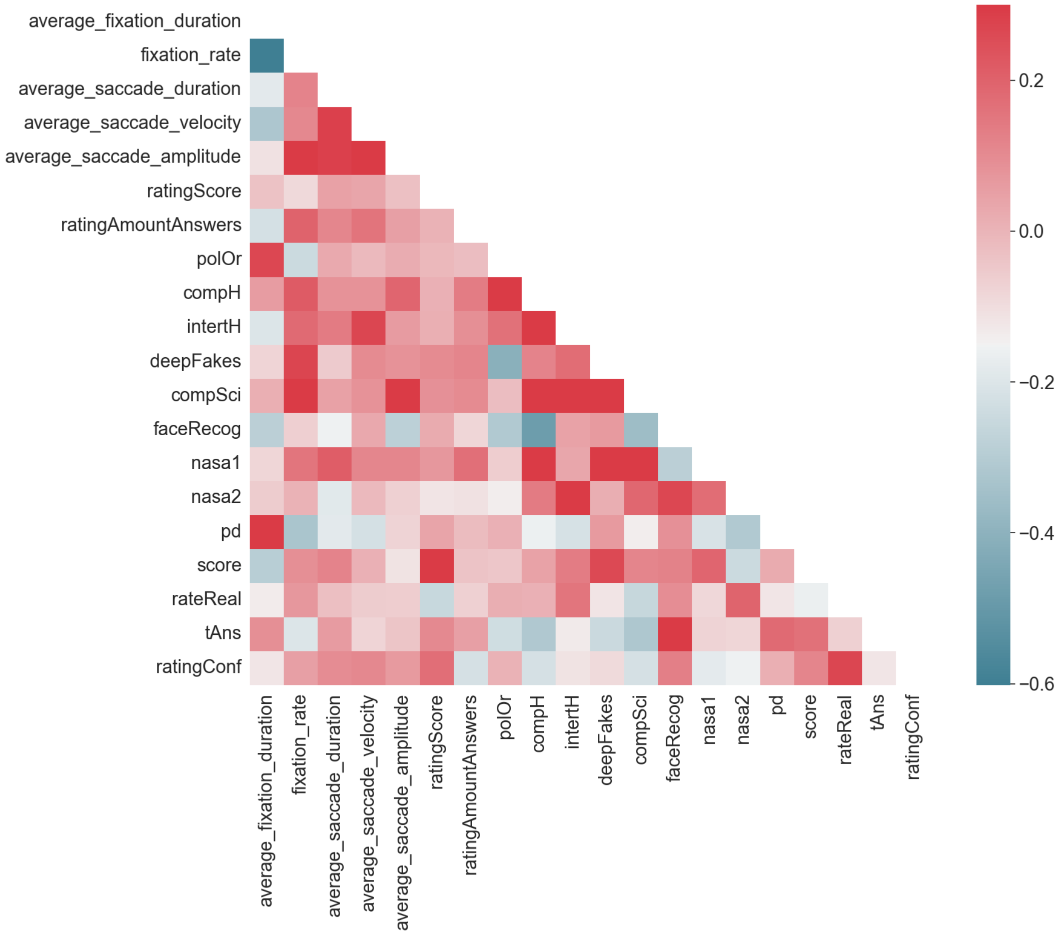


Figure 18. Heatmap of pearson-correlation coefficients.

Correlations related to experience with computer generated images:
daily computer use & daily internet use: ($r(20) = 0.5851938775882468$),
daily computer use & relation comp sci: ($r(20) = 0.5272388827859465$),
daily internet use & relation comp sci: ($r(20) = 0.39328865042452316$),
exp deep fakes & relation comp sci: ($r(20) = 0.45448908138361443$),

Correlation of eye movement measurements:
average saccade amplitude & average saccade velocity: ($r(20) = 0.5469757969442504$),
average fixation duration & fixation rate: ($r(20) = -0.6018253430083185$),
average saccade amplitude & fixation rate: ($r(20) = 0.30102470840722784$),
average fixation duration & average saccade velocity: ($r(20) = -0.3215678169500434$),
fixation rate & pupil dilation: ($r(20) = -0.32927651263186847$),

Correlation of experience with computer generated images and eye movement measurements:
nasa tlx 1 & relation comp sci: ($r(20) = 0.40774559713278513$),
daily computer use & nasa tlx 1: ($r(20) = 0.367328188336139$),
fixation rate & relation comp sci: ($r(20) = 0.34871351193143774$),
average saccade amplitude & relation comp sci: ($r(20) = 0.30659397809723055$),
exp deep fakes & nasa tlx 1: ($r(20) = 0.32332760125716764$)

Remaining correlations:
length text answer & relation comp sci: ($r(20) = -0.3193734743942356$),
daily computer use & length text answer: ($r(20) = -0.31388532858397283$),
face recognition & relation comp sci: ($r(20) = -0.3609175824105064$),
daily computer use & face recognition: ($r(20) = -0.48509502982364566$),
daily computer use & political orientation: ($r(20) = 0.39853307210655914$),
face recognition & length text answer: ($r(20) = 0.31237109825542914$),
rating evaluation & total score: ($r(20) = 0.3395631503817927$),
face recognition & political orientation: ($r(20) = -0.30955917962194546$),
exp deep fakes & political orientation: ($r(20) = -0.4114391143911438$),
nasa tlx 2 & pupil dilation: ($r(20) = -0.3128572418960731$),

3.3.2 Distribution of Rating Answers

When looking at the distribution of rating answers, we can investigate how a subject's behavior was influenced by what they believed about the image.

First, it is crucial to look at how the rating answers were distributed in general. As can be seen nicely in fig. 19, subjects tended to chose answers closer to either end of the scale. The neutral answer of 4 (neither real nor computer generated) was selected less than %5 percent of the time. Answers smaller than 4 mean subjects thought the image was real, while answers bigger than 4 mean the subject classified the image as computer generated.

Interesting to observe here is also the difference in slope on either end of the scale. While there seemed to be such a thing as ”I think this image might be computer generated” (rating score of 5), images were not as often rated real with low confidence (rating answer of 3).

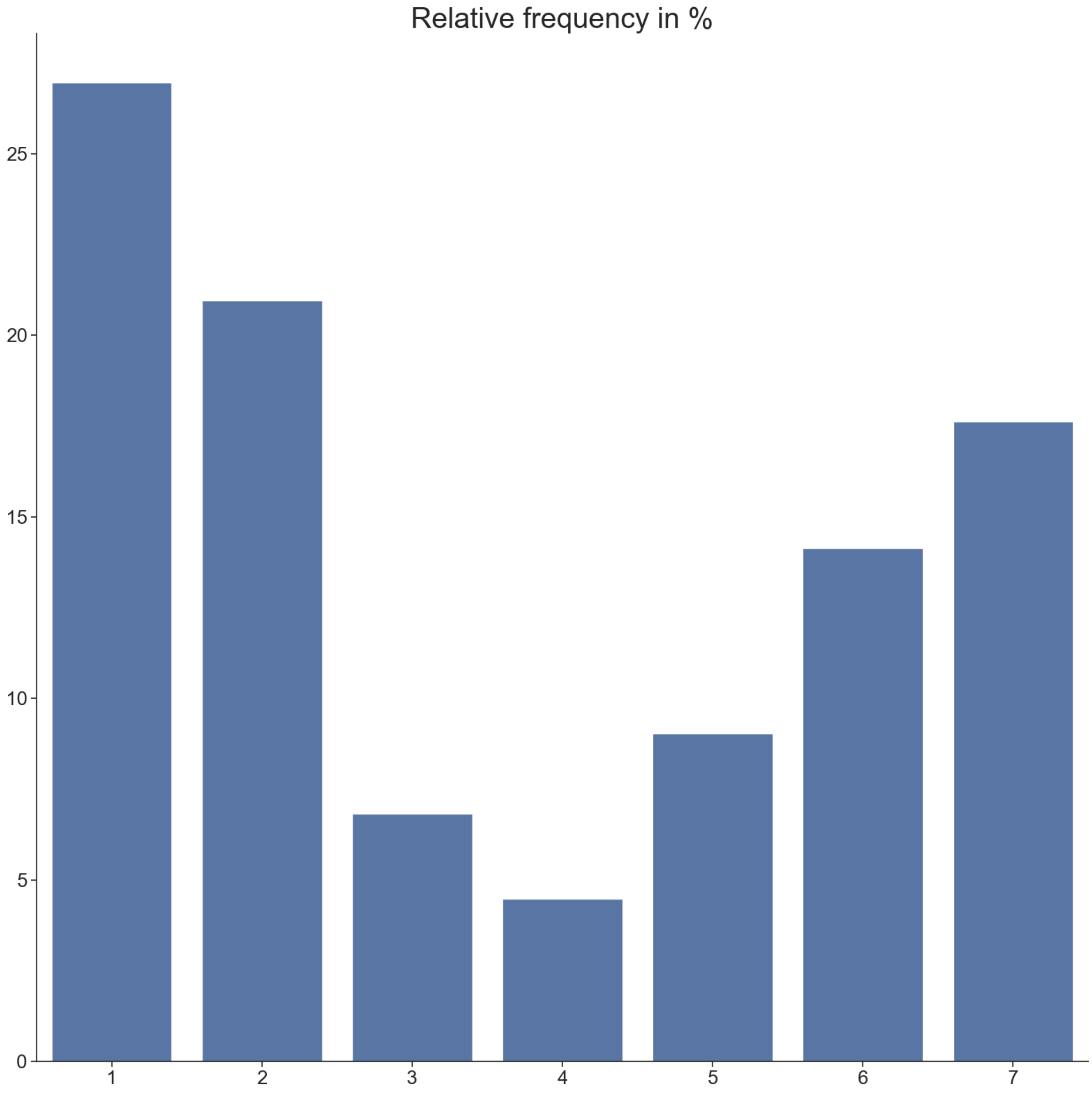
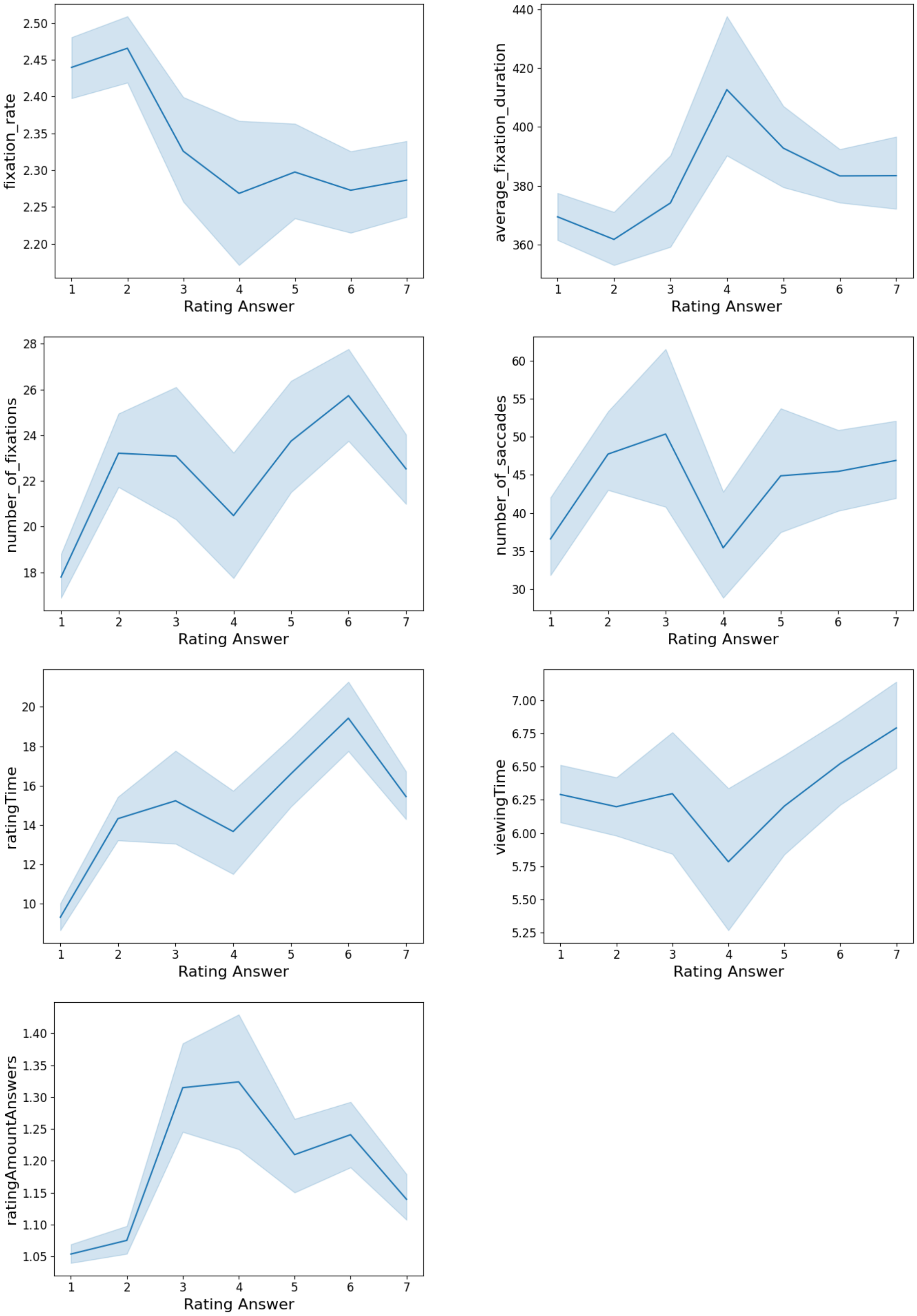


Figure 19. Relative frequency of rating answers.

The following plots depict (from top left to bottom right) the rating answers in relation to the fixation rate (1/s), average fixation duration (ms), number of fixations, number of saccades, the rating time (s) and the viewing time (s). Lastly, the amount of rating answers, which shows how many times the subjects selected an answer on the scale before continuing. A larger number of rating answers indicates multiple reevaluations, which could point to uncertainty).



fixation_rate
2.50
2.45
2.40
2.35
2.30
2.25
2.20
1
2
3
4
5
6
7
Rating Answer
average_fixation_duration
440
420
400
380
360
1
2
3
4
5
6
7
Rating Answer
number_of_fixations
28
26
24
22
20
18
1
2
3
4
5
6
7
Rating Answer
number_of_saccades
60
55
50
45
40
35
30
1
2
3
4
5
6
7
Rating Answer
ratingTime
20
18
16
14
12
10
1
2
3
4
5
6
7
Rating Answer
viewingTime
7.00
6.75
6.50
6.25
6.00
5.75
5.50
5.25
1
2
3
4
5
6
7
Rating Answer
ratingAmountAnswers
1.40
1.35
1.30
1.25
1.20
1.15
1.10
1.05
1
2
3
4
5
6
7
Rating Answer

3.3.3 NASA TLX

Looking at the resulting scores from the NASA TLX questionnaires can also give an indication to the subjects' cognitive load. The higher the score, the higher the estimated cognitive load. As we will compared cognitive load between viewing and rating phase, the results will give us an indication as to how much the task of each phase had impact on it.

The scores after the viewing phase were not significantly different between the groups ($t(20) = 1.86, p = .065$). Neither did the score after the rating phase differ($t(20) = -1.49, p = .137$).

Scores were significantly higher in the rating phase than in the viewing phase for both groups. Group 1: (V: ($M = 49.43, SD = 15.22$), R: ($M = 107.48, SD = 30.83$), ($t(20) = -17.15, p < 0.0001$)), group 2: (V: ($M = 45.12, SD = 15.52$)), R: ($M = 113.51, SD = 21.95$), ($t(20) = -26.67, p < 0.0001$)).

4 Discussion

4.1 Discussing our findings

4.1.1 Cognitive load and scores

By looking at differences **between the two groups**, we can investigate if subjects behaved differently depending on whether they were primed on the image condition (group 1) or not (group 2).

In the viewing phase, a lower fixation rate indicated a higher cognitive load in subjects from group 1. When viewing computer generated images, subjects that were primed on the image condition before had a higher fixation rate. It also seems like there was a higher variance in the average fixation duration for subjects from group 1, which should be further investigated. While long fixations indicate a higher cognitive load, a shorter duration indicates a lower cognitive load. This data could be analyzed in regard to subject's personal background.

These differences between the groups were not detectable in the rating phase, where comparing indicators of cognitive load did not indicate significant effects.

Concerning the classification and rating of images, group 1, that had seen the correct image conditions in the viewing phase, achieved a better overall score than group 2. This difference, we suspect, is mostly due to remembering image conditions from before. They did not remember and correctly identify every image though, as to be expected for 72 images. Interesting to see here is it seems there is a tendency for them to recall less realistic computer generated images (cg17) with a higher accuracy than more realistic ones (cg19). This could indicate they rely on recognizing obvious processing mistakes in the images more than the faces themselves. This theory could be supported by our finding that there was no significant correlation between subjects' self-assessed ability to recognize faces and their score. The latter effect was in accordance with paper [17], that also inquired this factor.

In addition to group 1's score, we can also see that they were answered with decreasing confidence for more computer generated images.

Looking at group 2's scores demonstrates that even the improvement that was made in computer generated images from 2017 to 2019 already decreased a human's ability to recognize them as fake substantially. While they detected images from cg17 more often than not, identifying images from cg18 already occurred with an equivalence to random guessing. They only were able to detect 50% of the images as computer generated. Images from cg19 were more often classified as real than as computer generated.

Interesting to highlight here is that in contrast to group 1, their confidence was not smaller when rating images from cg19 than it was for other image conditions. Meaning not only did they believe these computer images were real but they did so with fairly high confidence.

When comparing distributions **between conditions**, we can observe that for most measurements, the mean value of cg18 lies directly in between cg17 and cg19. This could suggest a continuous monotonous (maybe even linear) relationship between the measurements and the realness of computer generated images.

Effects **between the phases** can be looked at in two ways.
Firstly, when comparing cognitive load between phases, it is important to acknowledge there is likely already a significant difference due to the nature of the task. In the viewing phase, subjects were instructed to observe the images, while in the rating phase, they classified them it as real or computer generated. Comparing the NASA TLX results after the phases respectively supports this theory. On average, subject reported a significantly higher cognitive load in the rating phase.

Secondly, it is interesting to observe that this effect between phases was not found in every eye movement indicator for cognitive load. While saccade length and velocity[5] were consistently higher in the rating phase than they were in the viewing phase, thereby indicating higher cognitive load, this systematic difference was not found regarding the fixations. Only some groups for some image conditions showed a difference in fixation rate between the phases. This could be interpreted as not being due to the nature of the task alone but more due to seeing images for a second time.
Group 1 saw the images without knowing their conditions for the first time in the rating phase. When seeing images from cg19 for the second time, their fixation rate was

[5] It is useful to acknowledge these measurements are highly correlated [45]. This explains why their effects are so similar and is in accordance with our finding of their correlation.

significantly higher, indicating a lower cognitive load. Interpreting these results, one could infer their cognitive load was higher in the viewing phase due to the conflicting information subjects had to process. From looking at group 2 scores, we can confirm that more than half of the images from cg19 looked convincingly real. Group 1 seeing an image that looked real, even though they knew it was computer generated, could have resulted in cognitive conflict [44], which therefore increased cognitive load.

4.1.2 Personal background

Looking at correlations between subjects' personal background and their behavior in forms of cognitive load and rating can give us some insight on the interrelations between them. In this provisional analysis, we looked at eye movements combined for both phases. In future research, it would be interesting to see the phases analyzed separately.

We found that daily computer use, daily internet use, how much the subject's occupation was related to computer science and their familiarity with deep fakes were all heavily related. Subjects that reported higher values in these queries were deemed more experienced with the topic of computer generated images.

Subjects' experience seemed to have no effect on their score. This is partially not in accordance with findings from paper [17], which established a positive correlation between familiarity with deep fakes and subject's accuracy (corresponding to the score in our evaluation), but not on their confidence.

We found effects on cognitive load. But there was conflicting evidence: On one hand, subjects with more experience reported higher scores on the first NASA TLX and had a higher saccade amplitude, which both indicate a higher cognitive load. On the other hand, their overall fixation rate was also higher, which indicates a smaller cognitive load compared to subjects with less experience.

Eye movement measurements were correlated as expected. The saccade length and velocity were positively correlated. Fixation rate was negatively correlated with both the fixation duration pupil dilation. This is consistent with how each measurement relates to the cognitive load [45].

In our experiment, we deliberately gave subjects the option to take as much time as they needed for the viewing and the rating of images. Therefore, we were able to investigate differences in the time, as well as the number of fixations and saccades needed to observe and classify images from different conditions. Our findings indicate there were no significant differences here. Rating time and viewing time, as well as the number of fixations and saccades did not significantly differ for different image conditions.

While this indicates subjects observed real images with the same thoroughness and attention they looked the computer generated ones with, this might not be true in a real-world setting. Subjects in this experiment either knew from the start all images had the potential to be computer generated (group 1) or they likely caught on to this fact after a few trials (group 2). When consuming content on the internet, users might only detect images that already look computer generated or edited upon first glance as such. Thereby not evaluating all images by examining this level of detail, hypothetically reducing the time and number of fixations that is spent on an image with high levels of realness in a natural environment.

We furthermore recorded every subject's political orientation to see if people on the political right were worse at detecting false information [9] (in this case computer generated images). We found no correlation between political orientation and score. Subjects reported being further to the right of the political scale also did not have a higher realism rate, meaning they did not believe a higher percentage of the images were real compared to subjects who reported being part of the political left.

4.2 Practical Implications

Coming back to the topics that motivated this research, we will discuss potential practical implications of our findings.

Investigating what areas subjects utilize to detect computer generated images could help (1) educate internet users what distinct markers for synthetic information to look for when consuming content on the internet and (2) help improve the networks that can detect and flag this information.

Subjects viewed images with different levels of realness, which represented the improvement of this technology over the span of three years, from 2017 to 2019. We already saw that, while subjects were able to classify images from 2017 as fake more often than not, the opposite was true for images form 2019, while 2018 places somewhere in the middle. Since 2019, another three years have passed. If this trend continues, soon humans might have no change at detecting computer generated images.

The software that can be used to create this fake content is often open-source and available to everyone (StyleGAN2 for example is open source, everyone who understands how it works can use it. Even with no prior experience in computer science, websites like ”This person does not exist” allows users to create new faces by simply refreshing the page [24]). In the future, we should also focus on making networks that detect this content widely available. Therefore, it is necessary to shifting awareness from what is technologically possible to how technology can be applied in ways beneficial to humans. Either by (1) training networks to recognize false information. Or maybe interfere on a political level and passing laws that mandate adding markers in the form of noise to each peace of content GANs produce. This can be invisible to humans but clearly detectable by other networks, maybe using a principle similar to adversarial attacks (see [5]).

In regards to how humans perceive very realistic looking computer generated faces, our findings could imply that being alerted to their origin prior increases cognitive load. While one could conclude from the uncanny valley theory that objects with 100% human likeness evokes the highest affinity in the viewer, this might not be be the best goal for digital avatars. No matter how convincingly realistic they look, users will always have the background knowledge they are computer generated. If this knowledge alone evokes a higher cognitive load, the avatars could be a distraction to virtual working environments.

4.3 Further research

In our analysis, we have explored some effects and tendencies in our data. This is nowhere near exhaustive and with further analysis and investigation, more interesting findings could come from the data sets we produced in this study.

Furthermore, we propose some improvements that could be made to follow-up studies that are set up in a similar fashion.

4.3.1 Areas of Interest

Due to lack of time and high complexity, we did not investigate areas of interests (AOIs) in subjects' gaze patterns. We did though, in preparation of facilitating this analysis, compile a list of all fixations along with the coordinates of their centroids. By annotating the images based on the different areas (eyes, nose, mouth, hair, background, etc.) and comparing it to the list, AOIs can be detected.

After finalizing this analysis, it would be interesting to see:

- Which areas of the image did subjects focus on? How much time was spent on each percentually?
- Where were subjects' initial fixations? (When seeing the image for the first time in the viewing phase, where are the first couple of fixations? Are there differences between the groups; does knowing whether the image is real or fake change this behavior?)
- When seeing the image for the second time in the rating phase, are there different initial fixations? (differences between groups, will subjects jump straight to the area that helped them distinguish it in the viewing phase?)
- Where are the last fixations in the rating phase, what are the last areas that subjects look at before making their decision?
- Besides inspecting specific areas of the face, how much time was spent checking for symmetry? (for example, investigating how many times fixations changed from the left to the right side of the image)
- How does this gaze data relate to the text answers? Do subjects have a realistic assessment of where they were actually looking?
- Did the additional image condition we introduced (c3, open mouth - closed mouth) prompt a difference in gaze patterns? If the face's teeth were not visible, was less time spent looking at the mouth?

4.3.2 Potential Improvements

Since this is a bachelor's thesis, there are some things I have learned throughout this process. I dedicate this section to improvements I would make if I were to conduct another experiment following this one.

Due to better calibration results in the test runs, we decided to conduct our experiment in low-light conditions. While this increased the accuracy and validity of pupil detection, it also posed some challenges. As the room was only lit by the computer screen, we were concerned about subjects being discouraged from giving text answers due to not being able to see the keys on the keyboard sufficiently well. This would predominantly be an issue for people who spend less time on a computer and therefore presumably do not type blindly. We did find a negative correlation between subjects' daily computer use and the length of text answers they gave. This could potentially be either due to them being less comfortable with typing on a computer keyboard in general or due to subjects that need to look at the keys while typing not being able to see the keyboard sufficiently in low-light conditions. The latter should prompt a change that should be implemented in follow-up experiments.

While other papers also inquired subjects' daily computer use and did not find any positive effect on their scores (see [4]), we additionally inferred subjects' daily internet use. We suspected that, while being on a computer does not necessarily increase a person's likelihood of being exposed to deep fakes and computer generated content, spending time on the internet would. The high positive correlation we found daily computer and internet use suggests introducing this additional question did not necessarily explain more. Those personal factors might not be the best explanation why some subjects performed better than others. Therefore, in a follow-up experiment, we propose asking additional questions about (1) the subject's experience with image processing, especially software to manipulate images like Photoshop and (2) Not only internet use, but social media use in particular, as this might be the places people encounter deep fakes.

When conducting the experiment, we noticed subjects may need some time to get accustomed to the rating display. Comprehending the question, locating the scale and the text field took them a while when rating the first image. This assumption could be verified by seeing if the rating time was significantly higher than the rest of rating times for the first image in sequence for each participant. We therefore propose including a practice round, where subjects can test how the rating works before starting the actual rating, when data collection is already in progress.

During the experiment, some subjects gave verbal feedback that they were confused by the task. The instructions during the viewing phase were clear to them but when it came to filling out the NASA TLX, they were not sure what explicit task to base their answers on. This did not seem to pose a problem with their NASA TLX scores, as we still found sensible results but we still wanted to record this issue.

Working with the Tobii eye tracker and creating our own interface resulted in some issues that became apparent after the data collection. Starting the eye tracker simultaneously with the stimulus display prevented collecting excess data. But as the eye tracker took some time to detect the pupils every time it started back up, every recording having missed some samples at the beginning. This was only the case for around 150 samples (meaning at 600Hz, around 250ms). With an average fixation duration of around 300-400ms, we likely missed only one fixation. This occurred once at beginning of every phase, as well as after every break. Though this would probably not alter the bigger picture, it could be avoided next time by starting the eye tracker slightly in advance to displaying the stimuli.

Furthermore, including a blank screen for around 200ms before starting the eye tracker could aid in establishing a baseline of the subjects' pupil size. As measuring pupil dilation is mostly relevant relative to a person's individual baseline, applying the methods detailed in paper [19] will provide more meaningful results.

When analyzing the collected eye tracking data, we discovered there might be an issue with the validity data (see [33]) streamed from the eye tracker. Even when the pupil could not be detected properly and the eye tracker did not report valid data for gaze coordinates and pupil dilation, the validity indicators did not reflect this. If, in a future study, we wanted to investigate blinking patterns, having reliable validity data could facilitate this analysis. A blink can be found by detecting brief periods where validity is low due to both eyes being closed.

By eliminating certain images from the image dataset from paper [17], our goal was to reduce the relative impact the image's background had on the subjects' classification (see section 2.5). The word "background" appeared less in text answers than the word "hair" compared to text answers given in response to the original image data set. In our text answers, hair was mentioned 27.08%, compared to background with 20.42%. This suggests, though the background has lost some of its significance, it still plays an important role in discriminating between real and computer generated images. In order to further diminish this impact, the background could be removed completely in follow-up studies.

5 Conclusion

On the internet, we encounter an increasing amount of computer generated images and videos of faces. Investigating the consumers responses to this content is relevant to building technology that is beneficial to them.

In this study, we investigated how well subjects are able to discriminate between computer generated and real images of faces. Further, we looked at underlying cognitive processes. By tracking subjects' eye movements during the classification of images as real or computer generated, as well as when solely observing the images. When viewing the images, subjects were split into two groups, one of which saw the faces unbiased, while the other was primed on their origin. Based on this division, we could inquire how informing subjects about the image's origin influenced their behavior.

To conduct the experiment, we implemented an application, produced and analyzed data sets. This work provides an initial basis of experiments, discusses the findings of the acquired data and concludes some detailed directions for future research.

References

[1] Adrienne Goldstein for GMF. Social Media Engagement with Deceptive Sites Reached Record Highs in 2020. https://www.gmfus.org/news/social-media-engagement-deceptive-sites-reached-record-highs-2020, (2021, last accessed: 30.07.2022).

[2] Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., Nyström, M. One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior research methods*, 49(2):616–637, 2017. doi:10.3758/s13428-016-0738-9.

[3] Bundesregierung. Deepfakes: Ist das echt? https://www.bundesregierung.de/breg-de/themen/umgang-mit-desinformation/deep-fakes-1876736.

[4] Nicholas Caporusso, Kelei Zhang, and Gordon Carlson. Using eye-tracking to study the authenticity of images produced by generative adversarial networks. pages 1–6, 06 2020. doi:10.1109/ICECCE49384.2020.9179472.

[5] Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. In *International Conference on Learning Representations*, 2019. URL: https://openreview.net/forum?id=Syx_Ss05tm.

[6] Epic Games. MetaHuman. https://www.unrealengine.com/en-US/metahuman?sessionInvalidated=true, (2022, last accessed: 30.07.2022).

[7] D. Fallis. The epistemic threat of deepfakes. *Philos. Technol.*, 34:623–643, 2021. doi:10.1007/s13347-020-00419-2.

[8] Hong Gao, Efe Bozkir, Lisa Hasenbein, Jens-Uwe Hahn, Richard Göllner, and Enkelejda Kasneci. Digital transformations of classrooms in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3411764.3445596.

[9] R. Kelly Garrett and Robert M. Bond. Conservatives&#x2019; susceptibility to political misperceptions. *Science Advances*, 7(23):eabf1234, 2021. URL: https://www.science.org/doi/abs/10.1126/sciadv.abf1234, arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abf1234, doi:10.1126/sciadv.abf1234.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf, doi:10.48550/arXiv.1406.2661.

[11] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988. doi:10.1016/S0166-4115(08)62386-9.

[12] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 (2):65–70, 1979. URL: http://www.jstor.org/stable/4615733.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2017. URL: https://arxiv.org/abs/1710.10196, doi:10.48550/ARXIV.1710.10196.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. doi:10.1109/CVPR.2019.00453.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. doi:10.48550/arXiv.1912.04958.

[16] Thomas Kubler. The perception engineer's toolkit for eye-tracking data analysis. pages 1–4, 06 2020. doi:10.1145/3379156.3391366.

[17] Federica Lago, Cecilia Pasquini, Rainer Bohme, Helene Dumont, Valerie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, jan 2022. URL: https://doi.org/10.1109%2Fmsp.2021.3120982, doi:10.1109/msp.2021.3120982.

[18] Karanatsiou D. Vakali A. Lampridis, O. Manifesto: a human-centric explainable approach for fake news spreaders detection. *Computing 104*, page 717–739, 2022. doi:10.1007/s00607-021-01013-w.

[19] Fabius J. Van Heusden E. Van der Stigchel S. Mathôt, S. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior research methods*, 50(1):94–106, 2018. doi:10.3758/s13428-017-1007-2.

[20] Meta. Technologies that bring the world closer together. https://about.facebook.com/technologies/, (2022, last accessed: 30.07.2022).

[21] Microsoft. Microsoft Mesh. https://www.microsoft.com/en-us/mesh, (2022, last accessed: 30.07.2022).

[22] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 19(2):98–100, 2012. doi:10.1109/MRA.2012.2192811.

[23] Michoel L. Moshel, Amanda K. Robinson, Thomas A. Carlson, and Tijl Grootswagers. Are you for real? decoding realistic ai-generated faces from neural activity. *Vision Research*, 199:108079, 2022. URL: https://www.sciencedirect.com/science/article/pii/S0042698922000852, doi:https://doi.org/10.1016/j.visres.2022.108079.

[24] Phil Wang. This person does not exist. https://thispersondoesnotexist.com/, (2020, last accessed: 30.07.2022).

[25] PsychoPy. PsychoPy Home. https://www.psychopy.org/, (July 2022, last accessed: 25.07.2022).

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2939672.2939778.

[27] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research Applications*, ETRA ’00, page 71–78, New York, NY, USA, 2000. Association for Computing Machinery. doi:10.1145/355017.355028.

[28] Samantha Cole for Vice. Hacked News Channel and Deepfake of Zelenskyy Surrendering Is Causing Chaos Online. https://www.vice.com/en/article/93bmda/

hacked-news-channel-and-deepfake-of-zelenskyy-surrendering-is-causing-chaos-onli
(2022, last accessed: 30.07.2022).

[29] SR Labs. TOBII PRO SPECTRUM. https://www.srlabs.it/en/scientific-research/hardware-products/tobii-pro-spectrum/, (2022, last accessed: 28.07.2022).

[30] SR Research. Visual Angle Calculator. https://www.sr-research.com/visual-angle-calculator/, (2022, last accessed: 28.07.2022).

[31] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988. URL: https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1202_4, arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1202_4, doi:https://doi.org/10.1207/s15516709cog1202_4.

[32] Textblob. Textblob Documentation. https://textblob.readthedocs.io/en/dev/, (July 2022, last accessed: 26.07.2022).

[33] Tobii. What Do Validity Codes Mean? https://connect.tobiipro.com/s/article/What-Do-Validity-Codes-Mean?language=en_US#:~:text=The%20validity%20code%20ranges%20from,right%20eye%20by%20the%20system)., (2019, last accessed: 30.07.2022).

[34] Tobii. Tobii Pro Lab. https://www.tobiipro.com/product-listing/tobii-pro-lab/, (July 2022, last accessed: 25.07.2022).

[35] Tobii Website. Tobii IVT Fixation Filter. https://www.tobiipro.com/siteassets/tobii-pro/learn-and-support/analyze/how-do-we-classify-eye-movements/tobii-pro-i-vt-fixation-filter.pdf, (2012, last accessed: 25.07.2022).

[36] Tobii Website. Eye Tracker Data Quality Test Report. https://www.tobiipro.com/siteassets/tobii-pro/accuracy-and-precision-tests/tobii-pro-spectrum-accuracy-and-precision-test-report.pdf/?v=1.1, (June 2022, last accessed: 25.07.2022).

[37] Tobii Website. Eye tracking study recruitment – managing participants with vision irregularities. https://www.tobiipro.com/blog/eye-tracking-study-recruitment-managing-participants-with-vision-irregularities/, (June 2022, last accessed: 25.07.2022).

[38] Tobii Website. Glasses, Lenses and Eye Surgery. `https://help.tobii.com/hc/en-us/articles/210249865-Glasses-lenses-and-eye-surgery`, (June 2022, last accessed: 25.07.2022).

[39] Tobii Website. Tobii Integration with Eye Tracker Manager. `https://developer.tobiipro.com/eyetrackermanager/etm-sdk-integration.html`, (June 2022, last accessed: 25.07.2022).

[40] Tobii Website. Tobii Pro Lab User Manual, V. 1.194. `https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/Tobii-Pro-Lab-User-Manual/?v=1.194`, (June 2022, last accessed: 25.07.2022).

[41] Tobii Website. Tobii Pro Spectrum. `https://www.tobiipro.com/product-listing/tobii-pro-spectrum/`, (June 2022, last accessed: 25.07.2022).

[42] Tobii Website. Tobii Python Step-by-step guide. `https://developer.tobiipro.com/python/python-step-by-step-guide.html`, (June 2022, last accessed: 25.07.2022).

[43] Tom Cassauwers for Horizon. Can artificial intelligence help end fake news? `https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news`, (2019, last accessed: 30.07.2022).

[44] Patrick Weis and Eva Wiese. Cognitive conflict as possible origin of the uncanny valley. 10 2017.

[45] Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. Measuring cognitive load using eye tracking technology in visual computing. BELIV '16, page 78–85, New York, NY, USA, 2016. Association for Computing Machinery. `doi:10.1145/2993901.2993908`.

[46] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc., 2019. URL: `http://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf`.

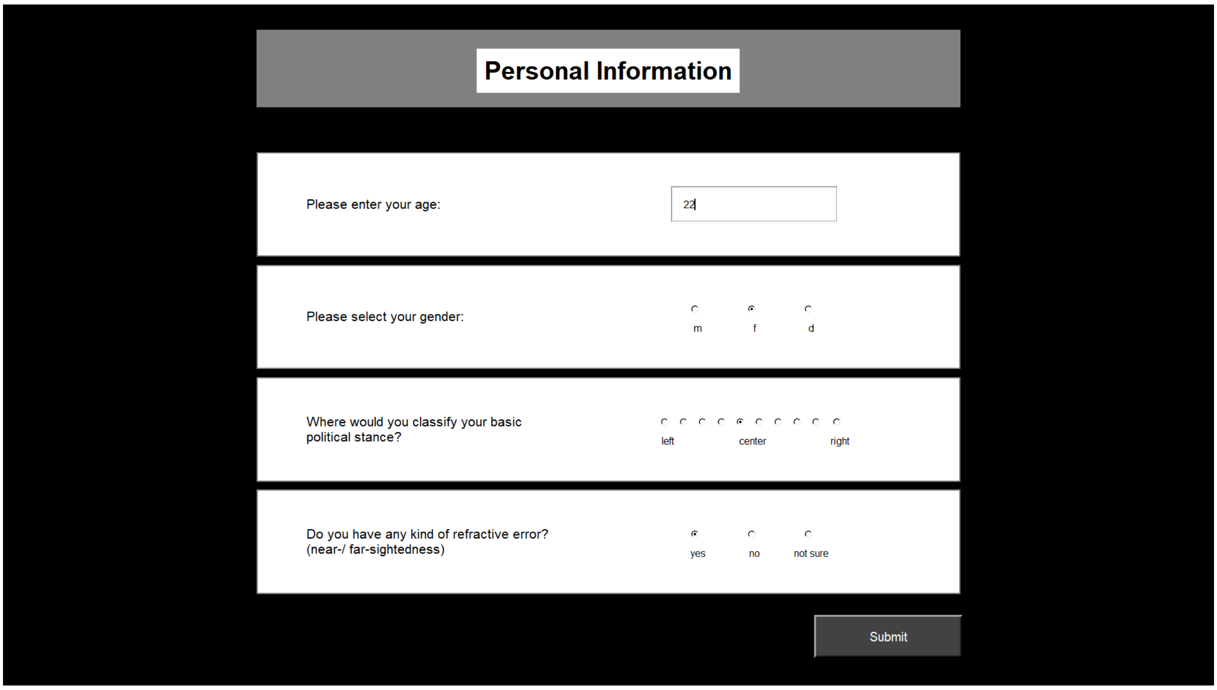
[47] Ömer Sümer, Efe Bozkir, Thomas Kübler, Sven Grüner, Sonja Utz, and Enkelejda Kasneci. Fakenewsperception: An eye movement dataset on the perceived believability of news stories. *Data in Brief*, 35:106909, 2021. URL: `https://www.`

sciencedirect.com/science/article/pii/S2352340921001931, doi:https://doi.org/10.1016/j.dib.2021.106909.

6 Appendix

6.1 Application UI

Screenshots that show the User Interface of the Application. Questionnaires in chronological order (Demographic (fig. 20), Eyesight (fig. 21), Familiarity (fig. 22)). Followed by NASA TLX (fig. 23) with weighting (fig. 24). Calibration (fig. 25). Viewing Screen with inactive (fig. 27) and active (fig. 28) "Continue" Button. Break Screen as an instance of "Instruction Screen" (fig. 26). Rating Screen (fig. 29) and Rating Screen with popup (fig. 30), in case there was no rating given.



Personal Information

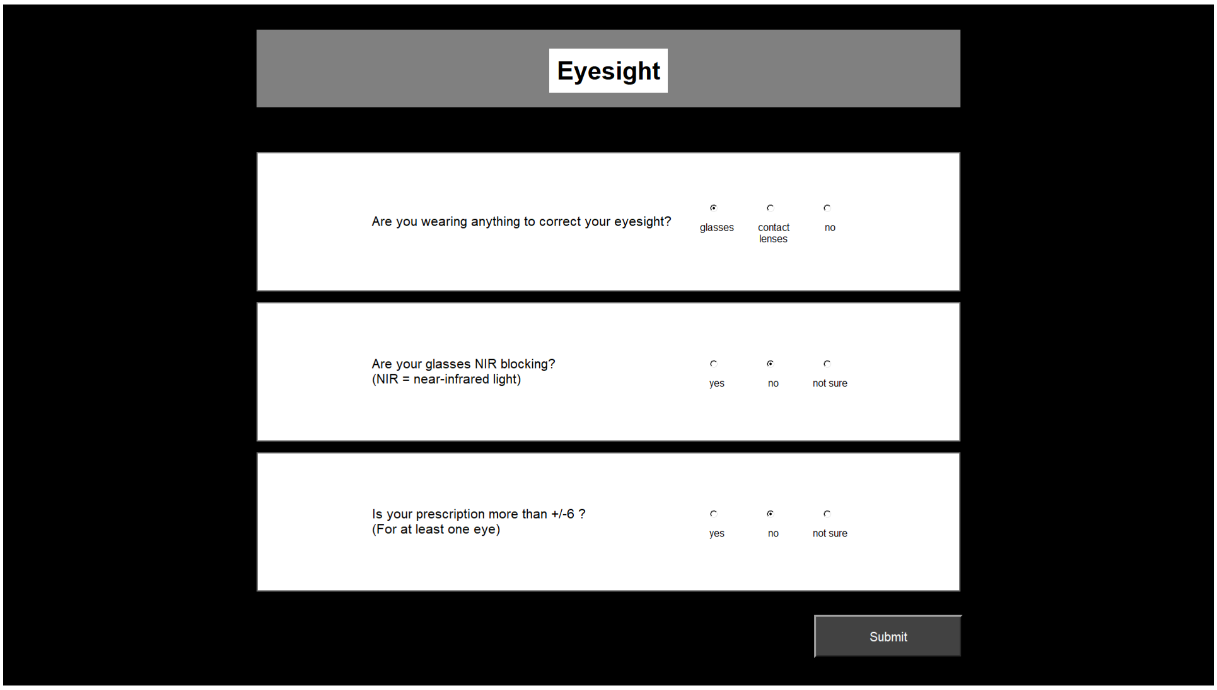
Please enter your age: 22

Please select your gender: ○ m ● f ○ d

Where would you classify your basic political stance? ○ ○ ○ ○ ● ○ ○ ○ ○ ○ (left – center – right)

Do you have any kind of refractive error? (near-/ far-sightedness) ● yes ○ no ○ not sure

Figure 20. Demographic Questionnaire.



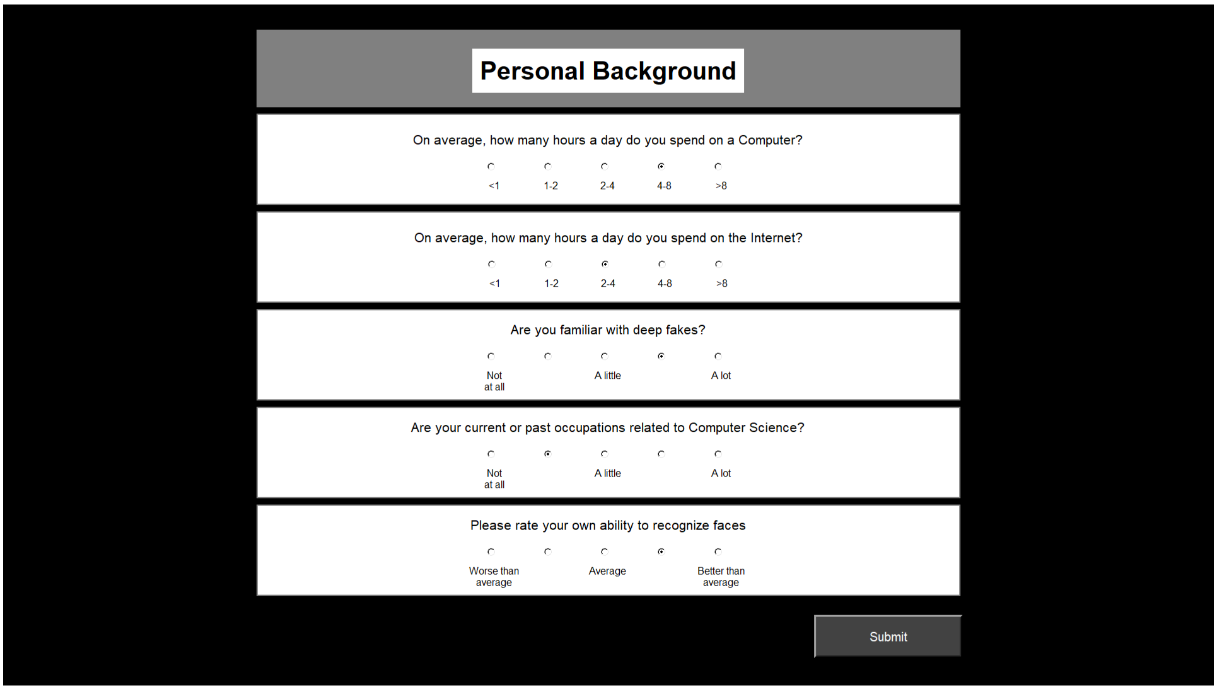
Eyesight

Are you wearing anything to correct your eyesight? ● glasses ○ contact lenses ○ no

Are your glasses NIR blocking?
(NIR = near-infrared light) ○ yes ● no ○ not sure

Is your prescription more than +/-6 ?
(For at least one eye) ○ yes ● no ○ not sure

Figure 21. Eyesight Questionnaire.



Personal Background

On average, how many hours a day do you spend on a Computer?
○ <1 ○ 1-2 ○ 2-4 ● 4-8 ○ >8

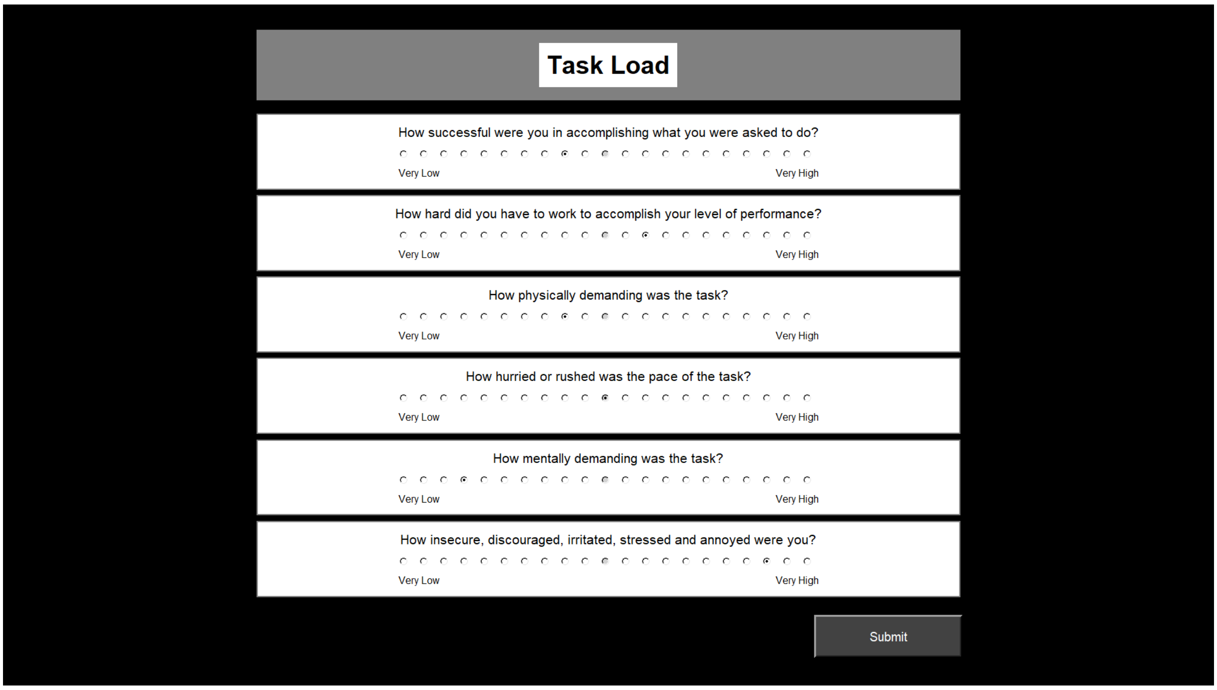
On average, how many hours a day do you spend on the Internet?
○ <1 ○ 1-2 ● 2-4 ○ 4-8 ○ >8

Are you familiar with deep fakes?
○ Not at all ○ ○ A little ● ○ A lot

Are your current or past occupations related to Computer Science?
○ Not at all ● ○ A little ○ ○ A lot

Please rate your own ability to recognize faces
○ Worse than average ○ ○ Average ● ○ Better than average

Figure 22. Familiarity Questionnaire.



Task Load

How successful were you in accomplishing what you were asked to do?

○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Very Low Very High

How hard did you have to work to accomplish your level of performance?

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○

Very Low Very High

How physically demanding was the task?

○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Very Low Very High

How hurried or rushed was the pace of the task?

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Very Low Very High

How mentally demanding was the task?

○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

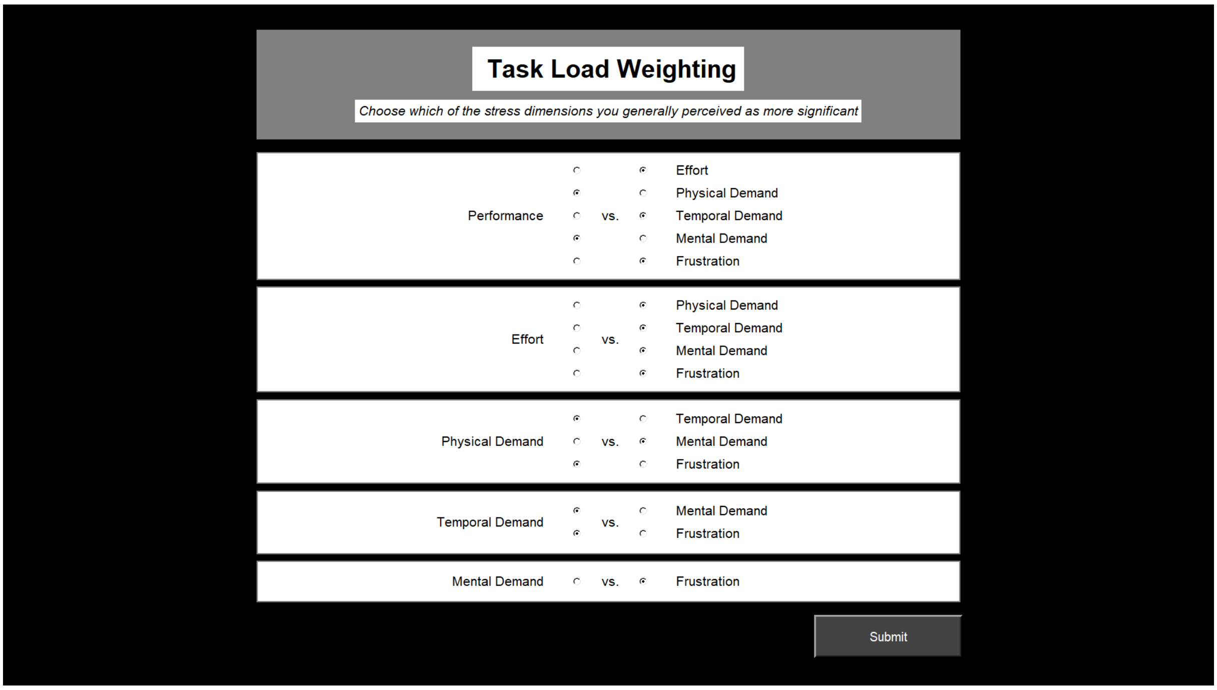
Very Low Very High

How insecure, discouraged, irritated, stressed and annoyed were you?

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○

Very Low Very High

Figure 23. NASA TLX.

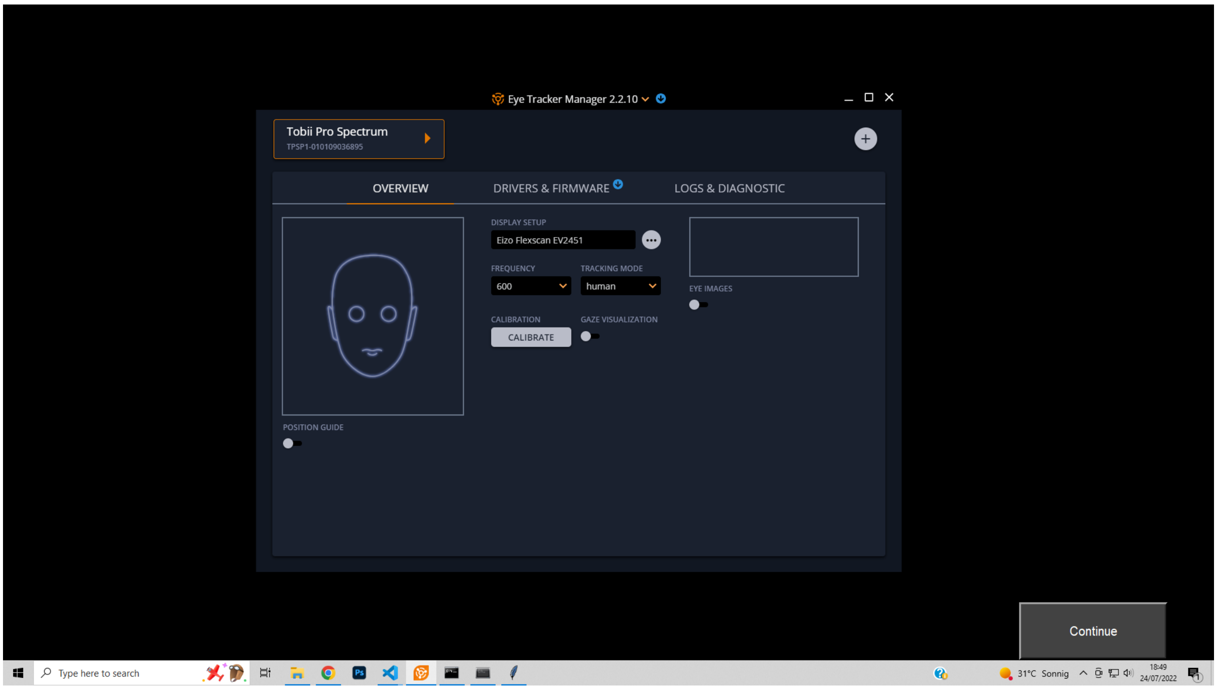


Task Load Weighting

Choose which of the stress dimensions you generally perceived as more significant

Performance	○	vs.	●	Effort
	●		○	Physical Demand
	○		●	Temporal Demand
	●		○	Mental Demand
	○		●	Frustration
Effort	○	vs.	●	Physical Demand
	○		●	Temporal Demand
	○		●	Mental Demand
	○		●	Frustration
Physical Demand	●	vs.	○	Temporal Demand
	○		●	Mental Demand
	●		○	Frustration
Temporal Demand	●	vs.	○	Mental Demand
	●		○	Frustration
Mental Demand	○	vs.	●	Frustration

Figure 24. NASA TLX Weighting.



Tobii Pro Spectrum
TPSP1-010109036895

DISPLAY SETUP: Eizo Flexscan EV2451

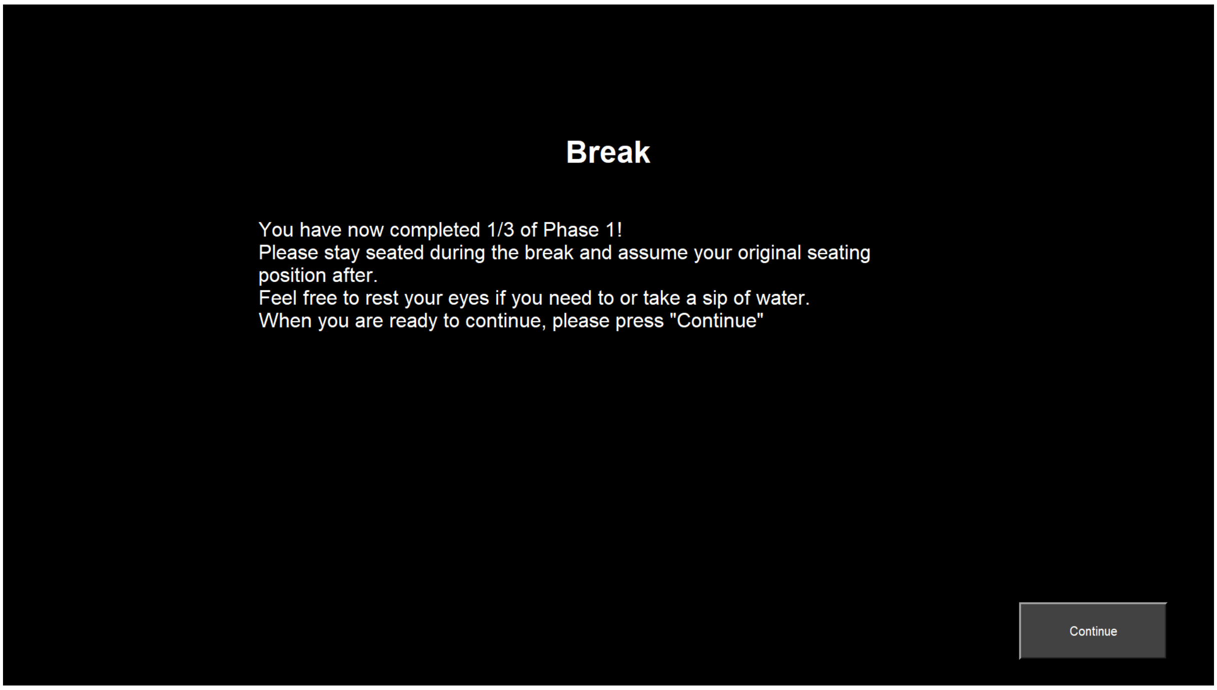
FREQUENCY: 600

TRACKING MODE: human

CALIBRATION

GAZE VISUALIZATION

Figure 25. Calibration using Tobii's Eye Tracker Manager.



Break

You have now completed 1/3 of Phase 1!
Please stay seated during the break and assume your original seating position after.
Feel free to rest your eyes if you need to or take a sip of water.
When you are ready to continue, please press "Continue"

Figure 26. Break Screen. Exemplary instance of "Instruction Screen"

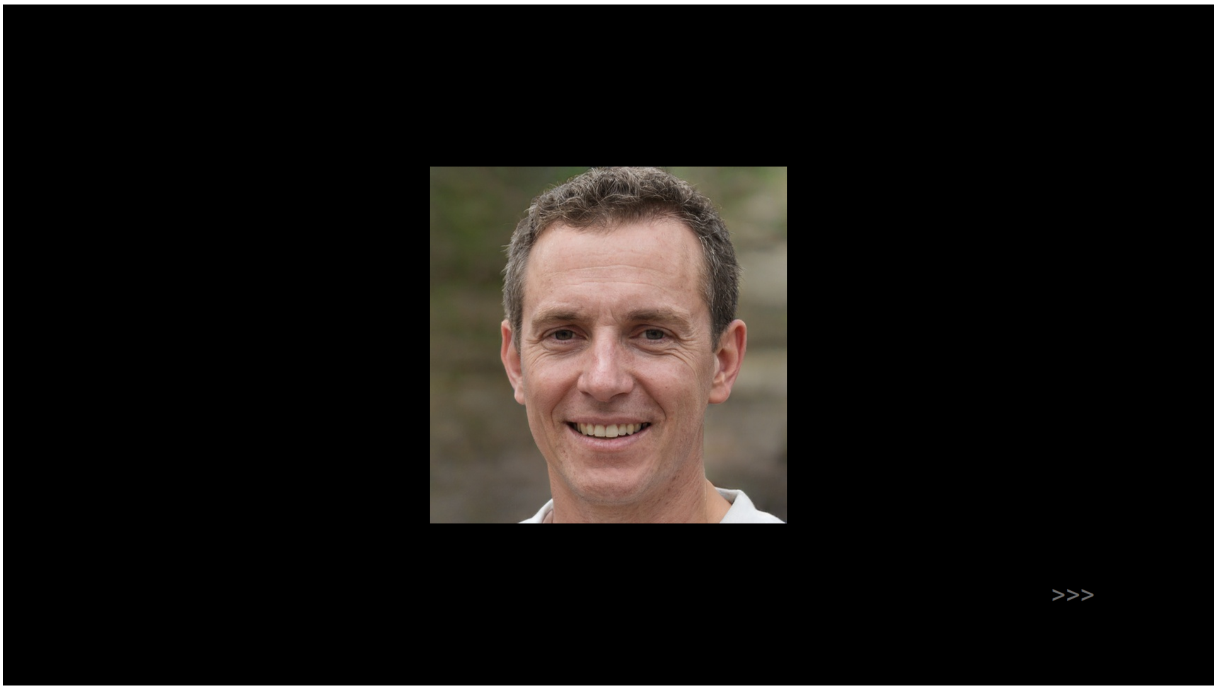


Figure 27. Viewing Screen with inactive "Continue"-button.

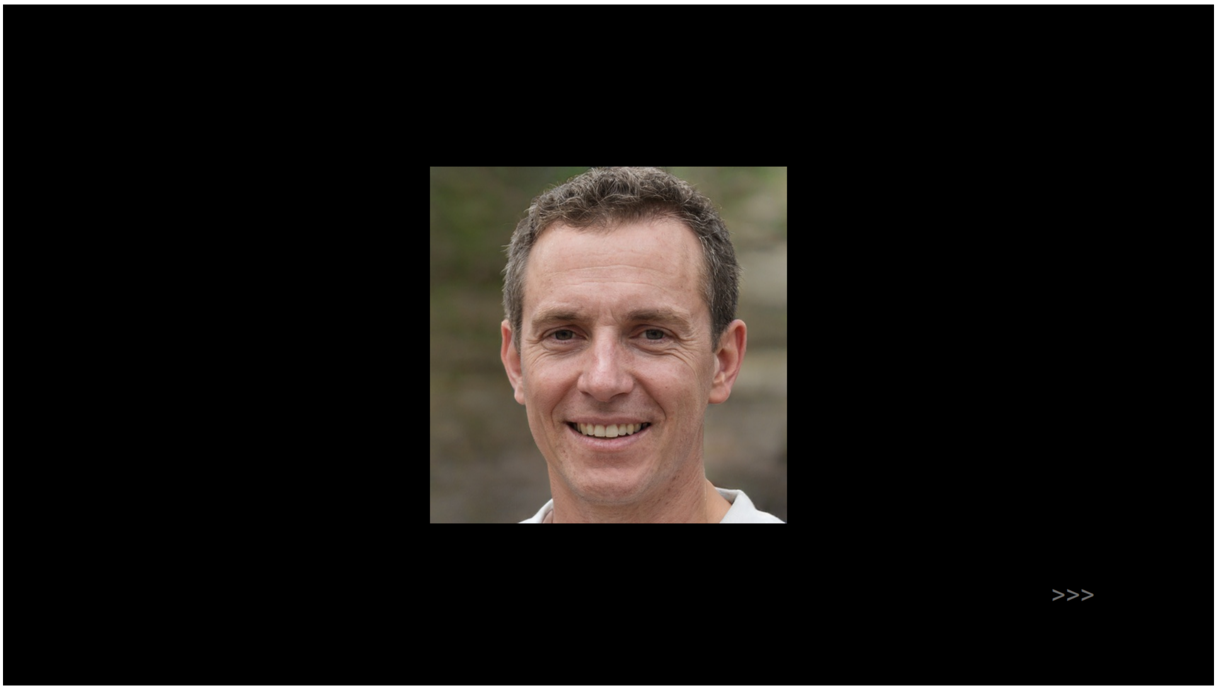
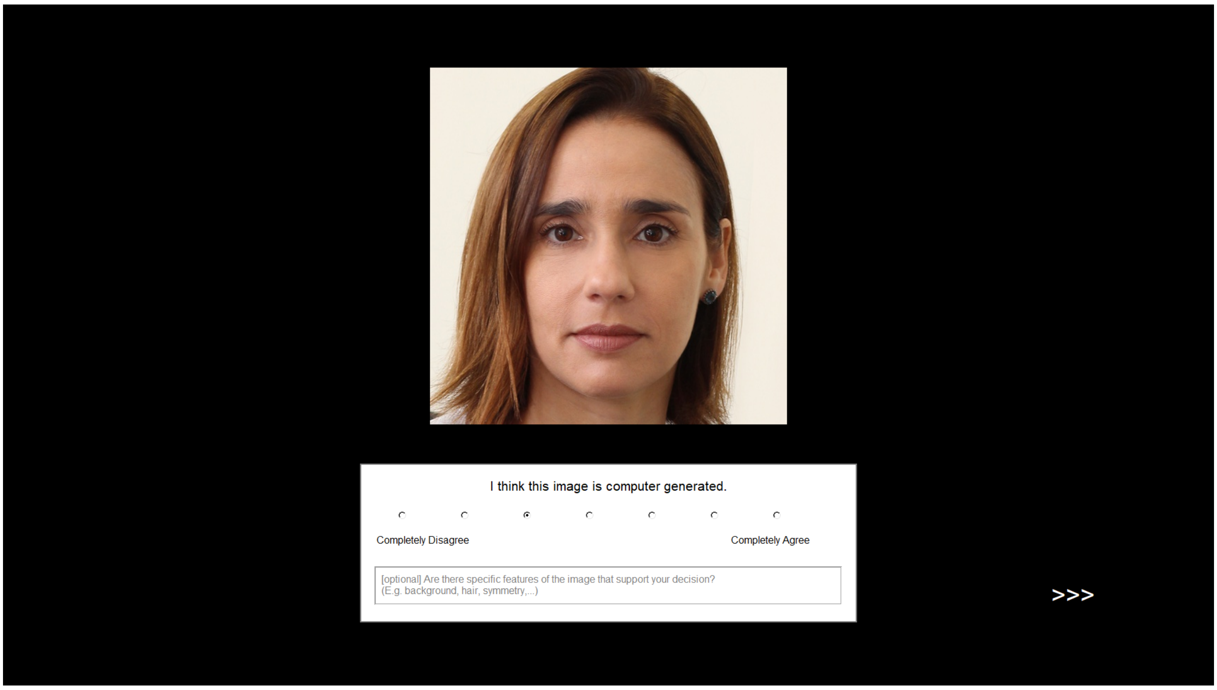


Figure 28. Viewing Screen with active "Continue"-button.



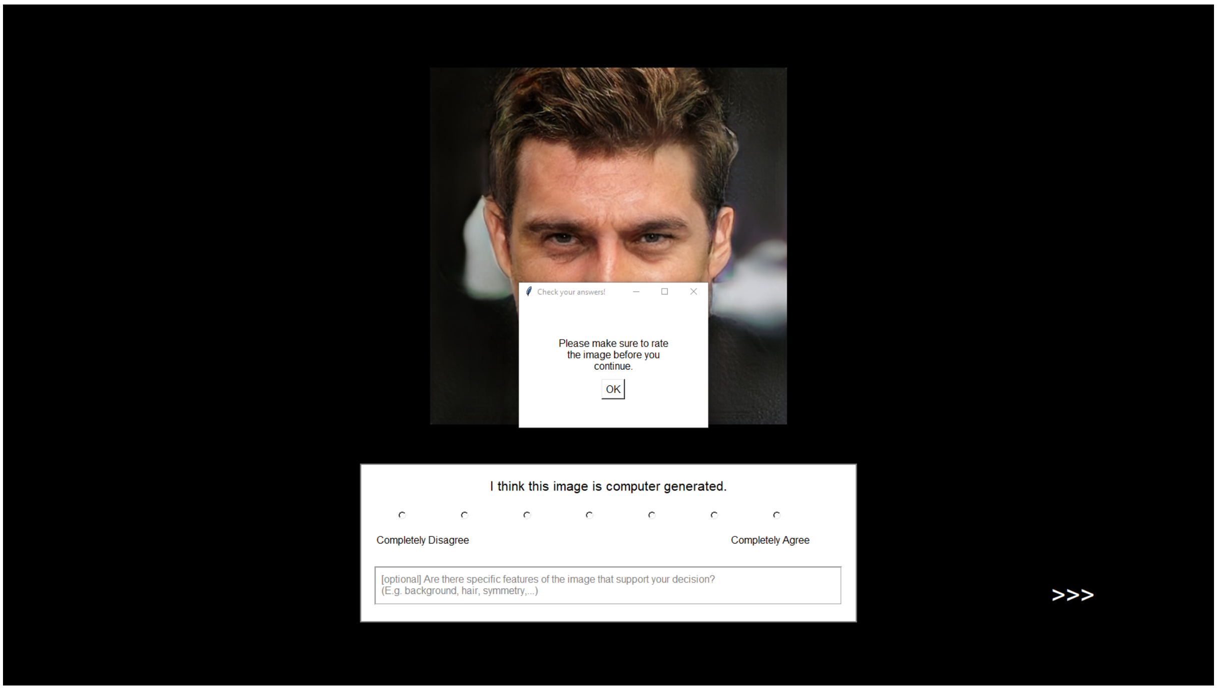
I think this image is computer generated.

○ ○ ● ○ ○ ○ ○

Completely Disagree Completely Agree

[optional] Are there specific features of the image that support your decision?
(E.g. background, hair, symmetry,...)

Figure 29. Rating Screen.



Check your answers!

Please make sure to rate the image before you continue.

I think this image is computer generated.

○ ○ ○ ○ ○ ○ ○

Completely Disagree Completely Agree

[optional] Are there specific features of the image that support your decision?
(E.g. background, hair, symmetry,...)

Figure 30. Rating Popup. Appeared when the user did not give a rating answer before pressing ”Continue”. Was also implemented in all questionnaires.

6.2 Additional plots

Plots that were not used in our analysis but may offer an additional angle to the visualization of our data.

6.2.1 Realism Rate

Similarly to the score (fig. 6), this plot shows the percentage of images that were rated as real by the subjects. In contrast to the score, this does not take into account the correctness of the rating. As the significant effects are analogous to the ones found when comparing the scores, this requires no further explanation.

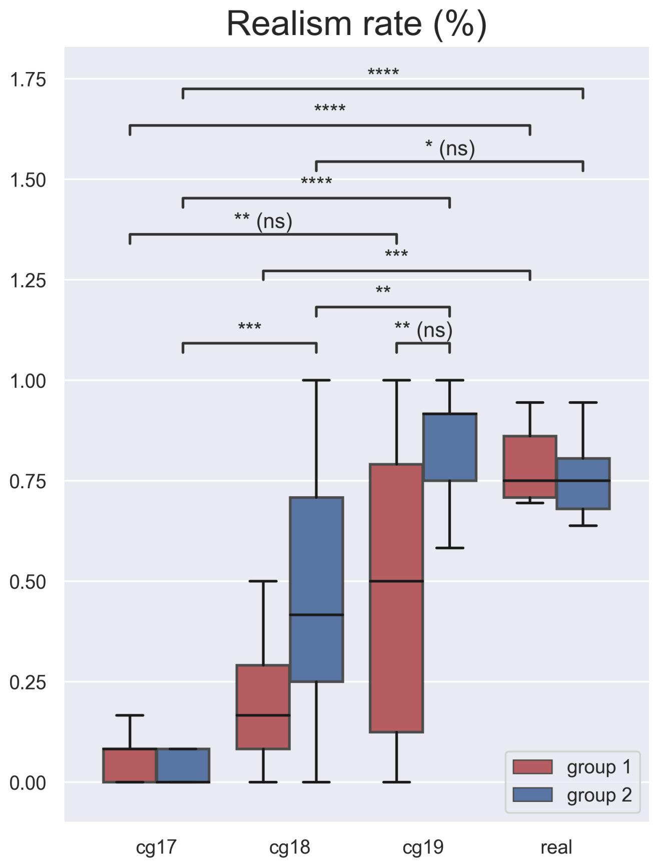


Figure 31. Realism rate; percentage of images subjects rated as real in each category. (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$)

Selbstständigkeitserklärung[6]

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

[6] Credit for the Design of the Title Page and this Declaration of Originality: Autonomous Vision Group, University of Tuebingen [https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/autonomous-vision/projects/bscmsc-theses/ (July 2022)]