Utrecht University

The Effect of Attitude Towards Computer Generated Faces on Face Perception

Eline Hoogers

6130011

Bachelor Thesis Artificial Intelligence

15 ECTS

Supervisor: Baptist Liefooghe

Second Assessor: Dominik Klein

July 2, 2021

Contents

Abstract

Attitudes towards computers have been becoming more negative over time. Earlier research has shown an unwillingness to cooperate with computers once their nature is revealed. Correspondingly, computer generated faces are perceived as less trustworthy than natural faces. This effect can be explained by either the synthetic appearance of the computer generated faces, or by a general bias towards artificial faces. Merely knowing that a face is computer generated could elicit a lower trustworthiness rating due to social categorization. Therefore, the aim of the current study was to examine whether the belief that a face is computer generated can impact how trustworthy and attractive it appears, using two experiment. In the first experiment, participants were presented two sets of natural faces and asked to rate these on trustworthiness and attractiveness. One of the sets was labeled computer generated and the other natural. In the second experiment, attitude towards computer generated faces was manipulated beforehand by presenting either a positive or a negative text before data collection. The results of both experiments show that merely labeling a face as computer generated is enough to make it appear less trustworthy but not less attractive. However, providing a positive or negative text before the task did not influence the ratings. It is implied that social categorization is involved in the discrimination between computer generated and natural faces.

Introduction

With technology evolving in a rapid tempo, artificial intelligence (AI) has an increasing role in our everyday life. While its products are applied in a considerable number of fields, it has the potential to become an even more essential part of life. However, the popularity of AI seems to be declining. Despite the increasing possibilities that AI entails, attitudes toward machines have become more negative over time (Gnambs & Appel, 2019). People seem to be cautious towards the use of robots, especially in the workplace. This trend is worrisome, because AI cannot fulfil its entire potential when people are not willing to embrace its appliance. Further development of this scientific field may not have the desired impact when its application is obstructed by these negative attitudes.

People seem to be hesitant to cooperate with a computer, but only when they are aware of the nature of the computer. A study by Ishowo-Oloko et al. (2019) showed the existence of a tradeoff between efficiency and transparency about the artificial nature of the computer. Participants were asked to engage in a prisoner's dilemma game, were they repeatedly had to make the decision to either cooperate with their partner or to defect. Results showed that bots were better at eliciting cooperation from its partners than humans. However, this efficiency decreased when the true nature of the bot was disclosed. These results demonstrate that an existing distrust towards bots minimizes their efficiency, even when the algorithm is highly developed. This lack of trust towards robots is present in multiple domains (Longoni et al., 2021; Promberger & Baron, 2006).

A possible solution for this problem is augmenting the computer with a high degree of anthropomorphism (Waytz et al., 2014). By increasing the human-likeness of a machine, people's trust is expected to improve. In this particular study, trust in an autonomous vehicle was higher when the vehicle was given a name, a gender and a voice. Given these findings, it would be

interesting to examine the effect of expanding a computer with a feature that could be seen as even more human-like: a face.

However, artificial faces are perceived as less trustworthy than natural faces. In a study by Balas and Pacella (2017), participants were asked to rate the trustworthiness of a set of faces. The set contained natural faces as well as faces with an artificial appearance. To create his second type of faces, photographs of human faces were manipulated to create an artificial looking image. Their study showed that the artificial faces were rated as less trustworthy than natural faces. Nonetheless, because of the synthetic appearance of the artificial faces, there was a clear distinction between the computer generated and the natural faces. On the one hand, this could have led to the artificial faces actually appearing less trustworthy. By manipulating the picture to make it seem artificial, the trustworthiness of the face could have been decreased. On the other hand, the results of Balas and Pacella could reflect a general attitude of participants towards AI, irrespective of differences in appearance between natural and artificial faces. Since preferences are thought to shape perception (Dunning & Balcetis, 2013), a negative attitude toward algorithms could impact the perception of computer generated faces. Hence, solely believing that a face is computer generated could affect how the face is perceived.

When seeing a face, people immediately form impressions about the person it belongs to (Zebrowitz & Montepare, 2008). In this process, all available information about the person can be used to create a useful impression. An important objective for creating this impression is the selection of a potential mate (Little, 2014). Faces that reflect beneficial qualities for a potential mate, are perceived as more attractive. The role of this evolutionary process in the perception of computer generated faces, however, remains insufficiently researched.

In creating an impression of a person, the visual information from the face as well as any

context that is provided can be used. For instance, the social group the person belongs to has an important impact on how that person is perceived. When someone belongs to the same social group, this person is likely to be perceived more positively (Ratner et al., 2014). This phenomenon is commonly known as social categorization. According to this theory, people identify themselves within certain social groups. The members of such groups are all similar in some aspect of their identity. For instance, people with the same ethnicity, age or occupation can consider each other as belonging to the same social group. When someone belongs to the same social group, this person is generally called an in-group member. When someone does not belong to the social group, this person is considered an out-group member.

This categorization into in- and out-groups is automatically instantiated upon seeing a face and can result in stereotyping attitudes and behaviors (Rule & Sutherland, 2017). Even when the attribution to an out-group is merely staged by the researchers, this phenomenon is found (Bernstein et al., 2007). In their study, participants were asked to complete a bogus personality test. They were then informed of having either a "green" or a "red" personality, while this was in fact not based on the actual outcome of the test. Afterwards they were presented photos of faces on a green or red background. They were told that this background denoted the personality type of the person presented. After observing all faces, they were again presented a series of faces. For each face they had to report whether they had seen it before. Their findings showed better recognition of the in-group faces than the out-group faces.

This theory could provide an explanation for the results of Balas and Pacella (2017). Artificial faces could be perceived as belonging to an out-group, whereas natural faces represent in-group members. Merely labeling a face as being computer generated could therefore elicit lower ratings. To further examine this theory, it is essential to manipulate the knowledge about the nature of the faces, without actually involving different types of faces.

The current study aims at testing whether the belief that a face is computer generated can impact its appearance. Contrary to earlier research, the effect of the synthetic appearance of the faces was excluded from the current study by only using natural faces. Hereby, purely the effect of bias was measured. Participants were presented two sets of natural faces. However, for one of these sets they were told that it was composed of computer generated faces. For each face, they were asked to rate how attractive and how trustworthy the face appeared.

Based on the research on the effect of preferences on perception, it was hypothesized that the faces labeled as computer generated would be rated less attractive than the faces with the natural label. As described earlier, facial attractiveness is thought to be related to mate selection (Little, 2014). Since computer generated faces do not represent potential mates, it is expected that participants have a preference for natural faces.

However, the research on social categorization suggests that creating an out-group can impact the impressions of a face on an even deeper level. Therefore, faces with the computer generated label are expected to appear as less trustworthy than faces with the natural label.

Experiment 1

Overview

The aim of the first experiment was to explore whether the perception of attractiveness and trustworthiness of faces labeled as computer generated differed from faces labeled as natural. To investigate this, an online survey was distributed in which participants were presented 2 sets of natural faces. Each set was comprised of 12 faces, which were selected based on their predetermined attractiveness and trustworthiness rating. To avoid dense results, the faces with the most extreme ratings on these attributes were selected. Therefore, each set contained 3 faces that were extremely attractive and extremely trustworthy, 3 faces that were extremely attractive and extremely untrustworthy, 3 faces that were extremely unattractive and extremely trustworthy and 3 faces that were extremely unattractive and extremely untrustworthy. For one of these sets participants were falsely informed that the faces were computer generated, whereas for the other set they were told that the faces were natural. They were asked to rate how attractive and how trustworthy each face appeared. To examine whether participants believed the labels, they were then asked to rate how believable they considered the computer generated faces.

It was hypothesized that the faces labeled as computer generated would be rated less trustworthy and less attractive than the faces labeled as natural. The experiment was preregistered on OSF (https://doi.org/10.17605/OSF.IO/CZYGP).

Method

Participants. A sample of 60 participants was recruited for the study through personal connections. Based on Brysbaert (2019) a minimum sample size of 50 participants was required for the study. Because of the counterbalancing, this number was increased to include 60 participants. Therefore, 15 participants were recruited for each of the four versions of the survey.

After outlier identification, one participant was excluded from data analysis and replaced by a new participant. The participants involved in the study were mostly female (48 female, 12 male). No restrictions on age were set. The age category of participants was measured instead of their exact age. The number of participants for each age category can be found in appendix A. The age group that was most represented was between 18 and 24 years old (27 participants). Participants were not provided any compensation.

Materials. A survey was created using Qualtrics online software (https://www.qualtrics.com). This survey consisted of 24 photos of natural faces, acquired from the Chicago Face Database (Ma et al., 2015). The faces were selected based on their trustworthiness and attractiveness rating as retrieved from the database. These ratings were measured on a 7-point Likert scale. The faces with the most extreme ratings on these characteristics were included in the survey. Thus, 6 faces were selected for each of the four possible extreme combinations. These four quadrants with their corresponding ranges of ratings were: trustworthy (4.8-5.5) and attractive (3.7–4.3), trustworthy (3.9–4.5) and unattractive (2.9–3.1), untrustworthy (2.0–2.5) and attractive (3.8–4.3), and untrustworthy (1.6–2.1) and unattractive (2.4-2.8). The codes corresponding to the faces that were included in each quadrant are provided in appendix B. To increase the credibility of the computer generated label, all faces were cropped in oval shape, excluding any hair and clothing. For each face, trustworthiness and attractiveness were rated on a 7-point Likert scale. Afterwards, believability of the faces labeled as computer generated was rated on a 7-point Likert scale. Participants used their own device to complete the survey.

Procedure. Participants were asked to complete a survey. The survey started with a cover story to prevent any suspicions that could elicit bias. The concept of computer generated faces was introduced and clarified in this story. Furthermore, there was described that earlier research had shown a difficulty in distinguishing computer generated from natural faces and that it was therefore important to examine how computer generated faces are perceived. The exact cover story can be retrieved from appendix C. Afterwards, informed consent was acquired and participants were asked for their age category and gender identification.

Hereafter, they were shown instructions about the task. In these instructions was mentioned whether the faces that would be shown in the following set were either computer generated or natural. After these instructions, 12 faces were presented one at a time. Participants were asked to rate each face on trustworthiness and attractiveness. The order in which the two characteristics were asked, was flipped among stimuli to retain alertness.

This task was repeated with a second set of faces, for which the instructions mentioned that the faces were of the opposite nature. Thus, if the first set of faces was labeled computer generated, the second set was labeled natural and vice versa. By means of counterbalancing, the order in which the labels were presented, differed between participants. Furthermore, it was counterbalanced among participants which set of faces was introduced by each label. After rating the 2 sets of faces, participants were asked to rate how believable they found the computer generated faces.

Data analysis. Only fully completed surveys were included in the data. The data was analyzed using SPSS. Outliers were identified and removed when they deviated more than 2.5 SD from the group mean. This resulted in the removal of one outlier. This participant was replaced by a new participant, after which the outlier identification procedure showed no more

outliers.

Two repeated measures ANOVA's were then executed with three within-subject factors: Label, Attractiveness and Trustworthiness. Label refers to whether the face was classified as natural or computer generated. The factors Attractiveness and Trustworthiness refer to the attractiveness and trustworthiness rating that was supplied in de database for each face. For each attribute, the rating was classified as either low or high, corresponding to the four quadrants (See Appendix B). In the first analysis, the trustworthiness rating was included as the dependent variable. In the second analysis, the attractiveness rating was the dependent variable.

Afterwards, two correlational analyses were performed: believability and difference in trustworthiness, and believability and difference in attractiveness. The difference in trustworthiness was determined by subtracting the trustworthiness rating of the faces labeled as computer generated from that of the faces with the natural label. For attractiveness, the same procedure was used.

Results

Trustworthiness. The results of the repeated measures ANOVA for trustworthiness are summarised in Table 1. The descriptive statistics for each main effect are displayed in Table 2.

As hypothesized, a significant main effect of Label on Trustworthiness was found. Faces labeled as computer generated were rated as less trustworthy than the faces with the natural label.

Furthermore, significant main effects were found for the pre-determined attractiveness and trustworthiness ratings. No significant interaction-effects were found.

Table 1

Repeated Measures ANOVA Results for Trustworthiness in Experiment 1

Within-subjects effects	$F(1,59)$	p	η^2
Label	5.58	.022	.09
Attractiveness	215.96	< .001	.79
Trustworthiness	69.09	< .001	.54
Label x Attractiveness	0.57	.454	.01
Label x Trustworthiness	0.59	.445	.01
Attractiveness x Trustworthiness	2.99	.089	.05
Label x Attractiveness x Trustworthiness	0.19	.667	.00

Attractiveness. The results of the repeated measures ANOVA for attractiveness are summarised in Table 3. The descriptive statistics for each main effect are displayed in Table 2. Contrary to our hypotheses, no significant main effect of Label on Attractiveness was found. The faces labeled computer generated were rated less attractive than the faces with the natural label, but this effect was not significant.

Furthermore, significant main effects were found for the pre-determined attractiveness and trustworthiness ratings. A significant interaction-effect between Attractiveness and Trustworthiness was found. However, no interaction-effects were found for Label and Attractiveness, Label and Trustworthiness, and Label, Attractiveness and Trustworthiness.

Table 2

Means and Standard Deviations of Trustworthiness and Attractiveness in Experiment 1

Within-subjects factor	Trustworthiness		Attractiveness	
	M	*SD*	*M*	*SD*
Label				
Natural	3.93	0.67	3.93	0.67
Computer generated	3.78	0.70	2.98	0.70
Attractiveness				
Low	2.13	0.71	2.13	0.71
High	3.93	0.83	3.93	0.83
Trustworthiness				
Low	3.38	0.72	2.63	0.70
High	4.33	0.63	3.42	0.74

Believability. No significant correlation was found between Believability and the difference in trustworthiness rating between the two labels (r (59) = -.03, p = .849). For attractiveness, also no significant correlation was found (r (59) = .01, p = 0.948).

Table 3

Repeated Measures ANOVA Results for Attractiveness in Experiment 1

Within-subjects effects	$F(1,59)$	p	η^2
Label	1.61	.210	.03
Attractiveness	174.34	< .001	.75
Trustworthiness	337.57	< .001	.85
Label x Attractiveness	0.02	.899	.00
Label x Trustworthiness	0.05	.822	.00
Attractiveness x Trustworthiness	25.01	.000	.30
Label x Attractiveness x Trustworthiness	0.04	.844	.00

Discussion

Based on the study by Balas and Pacella (2017), it was predicted that faces labeled as computer generated would be rated as less trustworthy due to social categorization (Bernstein et al., 2007). Furthermore, the faces labeled as computer generated were expected to be rated less attractive than the natural faces, because of evolutionary motives (Little, 2014). However, these predictions were only partly supported by our findings. The predicted effect of the labeling was found on the trustworthiness rating but not on the attractiveness rating. These findings suggest that merely the belief that a face is computer generated is enough to impact how trustworthy it appears, but not how attractive is appears.

These findings in part support the hypothesis that the perception of computer generated faces is affected by bias due to social categorization. To further examine the effect of bias on face perception, a second experiment was conducted. In this experiment, the bias towards computer

generated faces was manipulated prior to data collection by presenting either a positive or a negative text about computer generated faces.

By manipulating the bias prior to data collection, the effects of social categorization are expected to increase. This effect has been demonstrated in a study by Derks et al. (2014). In their study, Muslim participants were asked to describe an incident in which they had experienced discrimination based on their religion or respect because of their religion. Thereby, the amount of social identity threat was manipulated. Afterwards, they performed an evaluative priming task with Muslim faces and non-Muslim faces as primes. Their results show a stronger differentiation between groups when there was a high threat to personal identity. These findings suggest that in the current study, the effect of bias on perception could be amplified by manipulating the attitude towards computer generated faces.

It is therefore expected that the effect of labeling on the trustworthiness rating, as found in the first experiment, is larger in the negative condition than in the positive condition. Thus, an interaction effect between the label and condition is expected for trustworthiness. Since no effect was found on the attractiveness rating, this effect is expected to remain absent in the second experiment.

Experiment 2

Overview

To further investigate the impact of labeling on the perceived trustworthiness, a follow-up experiment was conducted. In this experiment, the method from the first experiment was adopted with an addition. Prior to the experiment, participants read either a positive or a negative text about the adoption of computer generated faces. By presenting this text, the effect of social categorization was expected to be increased due to social identity threat (Derks et al., 2014).

To assess the effect of the manipulation story on the attitude towards computer generated faces, participants were asked to rate their general attitude towards computer generated faces at the end of the survey.

It was hypothesized that faces labeled as computer generated would be rated less trustworthy but not less attractive than faces with the natural label. Furthermore, the effect of labeling on trustworthiness was expected to be larger in the positive condition than in the negative condition. The experiment was preregistered on OSF (https://doi.org/10.17605/OSF.IO/CZYGP).

Method

Participants. A sample of 130 participants was recruited for the follow-up experiment through social media. After outlier identification, 6 participants were excluded from data analysis. Therefore, the final sample consisted of 124 participants. 61 participants received the positive manipulation story and 64 participants received the negative story. The participants involved in the study were mostly female (96 female, 27 male, 1 non-binary). The age group that was most represented was between 18 and 24 years old (see appendix D). Among the participants, 1 gift voucher of 25 euro was allotted.

Materials and procedure. The materials and procedure of the experiment were similar to those used in experiment 1. However, the cover story was adapted to include a manipulation of attitude towards computer generated faces. The participants either receive a positive or a negative manipulation story. The beginning of both stories was equal to the original cover story from experiment 1. The concept of computer generated faces was introduced and participants were informed that it was very difficult for people to distinguish computer generated from natural faces. The second part of the story differed between the positive and negative condition. In the positive story, the alluring possibilities of this technology were highlighted. In the negative story, the threatening risks of this technology were highlighted. The exact stories are provided in appendix E. Subsequently, it was checked whether the participant had already participated in the first experiment. These participants were excluded from further participation in the follow-up experiment. To measure the impact of the manipulation stories, participants were asked to rate their general attitude towards computer generated faces on a 7-point Likert scale at the end of the survey.

Data analysis. Only fully completed surveys were included in the data. The data was analyzed using SPSS. Outliers were identified and removed when they deviated more than 2.5 SD from the group mean. This resulted in the removal of 6 outliers.

Two repeated measures ANOVA's were then executed to analyze the effects of three within-subjects factors (Label, Attractiveness and Trustworthiness) and one between-subjects factor (Condition). The factor Label corresponded to the classification of the face as either natural or computer generated. The factors Attractiveness and Trustworthiness referred to the pre-existing attractiveness and trustworthiness rating of each face in the database. These factors were

each split up in two levels: Low and High (See Appendix B). The Condition of each participant was determined by which manipulation story they had received. This was either the positive of the negative story. The ANOVA's were executed with trustworthiness as the dependent variable and with attractiveness as the dependent variable.

Lastly, four correlational analyses were performed: attitude and difference in trustworthiness, attitude and difference in attractiveness, believability and difference in trustworthiness, and believability and difference in attractiveness.

Results

Trustworthiness. The results of the repeated measures ANOVA for trustworthiness are displayed in Table 4. The corresponding means and standard deviations can be found in Table 5.

As hypothesized, a significant main effect of Label on the trustworthiness rating was found. Faces labeled as computer generated were rated as less trustworthy than the faces with the natural label. Moreover, significant main effects were found for the pre-existing trustworthiness and attractiveness ratings. No significant main effect of Condition was found

Contrary to our hypotheses, no significant interaction-effect was found between Label and Condition. Significant interaction-effect were found for Label and Trustworthiness. No further interaction-effects were found.

Table 4

Repeated Measures ANOVA Results for Trustworthiness in Experiment 2

Factors	$F(1,122)$	p	η^2
Label	6.46	.012	.05
Attractiveness	67.85	< .001	0.36
Trustworthiness	350.73	< .001	.74
Condition	0.01	.930	< .01
Label x Attractiveness	0.79	.377	.01
Label x Trustworthiness	4.29	.041	.03
Label x Condition	0.88	.350	.01
Attractiveness x Trustworthiness	3.23	.075	.03
Attractiveness x Condition	1.54	.216	.01
Trustworthiness x Condition	< 0.01	.990	< .01
Label x Attractiveness x Trustworthiness	0.06	.841	< .01
Label x Attractiveness x Condition	2.10	.150	.02
Label x Trustworthiness x Condition	1.13	.289	.01
Attractiveness x Trustworthiness x Condition	0.17	.684	< .01
Label x Attractiveness x Trustworthiness x Condition	1.14	.287	.01

Table 5

Means and Standard Deviations of Trustworthiness in Experiment 2

Within-subjects factor	Positive		Negative		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*
Label						
Natural	4.13	0.57	4.18	0.68	4.15	0.63
Computer generated	4.02	0.73	3.95	0.72	3.98	0.72
Attractiveness						
Low	3.86	0.69	3.91	0.67	3.88	0.68
High	4.28	0.57	4.22	0.53	4.25	0.55
Trustworthiness						
Low	3.59	0.66	3.58	0.61	3.59	0.63
High	4.55	0.66	4.54	0.61	4.55	0.63

Attractiveness. The results of the repeated measures ANOVA are shown in Table 6. The means and standard deviations are displayed in Table 7.

As hypothesized, no significant main effect of the label on the attractiveness rating was found. Furthermore, no significant main effect of Condition was found. Significant main effects of Attractiveness and Trustworthiness were found.

No significant interaction-effect between Label and Condition was found. However, a significant interaction-effect between Attractiveness and Trustworthiness was found. No further interaction-effects were found.

Table 6

Repeated Measures ANOVA Results for Attractiveness in Experiment 2

Factors	$F(1,122)$	p	η^2
Label	2.78	.098	.02
Attractiveness	548.14	< .001	.82
Trustworthiness	321.04	< .001	.73
Condition	0.32	.631	< .01
Label x Attractiveness	0.49	.484	< .01
Label x Trustworthiness	0.34	.563	< .01
Label x Condition	0.46	.497	< .01
Attractiveness x Trustworthiness	31.28	< .001	.20
Attractiveness x Condition	0.12	.727	< .01
Trustworthiness x Condition	1.67	.198	.01
Label x Attractiveness x Trustworthiness	0.62	.432	.01
Label x Attractiveness x Condition	1.71	.194	.01
Label x Trustworthiness x Condition	0.25	.620	< .01
Attractiveness x Trustworthiness x Condition	0.01	.944	< .01
Label x Attractiveness x Trustworthiness x Condition	0.17	.682	< .01

Table 7

Means and Standard Deviations of Attractiveness in Experiment 2

Within-subjects factor	Positive		Negative		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*
Label						
Natural	3.31	0.71	3.34	0.69	3.32	0.69
Computer generated	3.21	30.70	3.29	0.73	3.25	0.71
Attractiveness						
Low	2.41	0.78	2.49	0.77	2.45	0.77
High	4.11	0.78	3.29	1.16	4.13	0.77
Trustworthiness						
Low	2.94	0.66	2.95	0.66	2.95	0.66
High	3.57	0.75	3.68	0.70	3.63	0.73

Attitude. The main effect of the condition on the attitude towards computer generated face was almost significant ($F(1, 122) = 3.50$, $p = .064$, $\eta^2 = .028$). Participants who received the negative manipulation story rated their attitude as lower ($M = 3.86$, $SD = 1.24$) than participants who received the positive manipulation story ($M = 4.28$, $SD = 1.27$).

Furthermore, Attitude did not correlate significantly with the difference in trustworthiness rating (r (123) = -.12, $p = .205$) and attractiveness rating (r (123) = .05, $p = .565$) between the two labels.

Believability. The correlation between Believability and the difference in trustworthiness rating between the two labels was almost significant (r (123) = -.17, p = .068). For attractiveness, a significant negative correlation was found (r (123) = -.20, p = 0.028).

Discussion

The current findings suggest that bias against computer generated faces is enough to make a face appear less trustworthy but not less attractive. However, this effect was not increased by manipulating the bias with either a positive or negative story. The manipulation did have an almost significant effect on the attitude towards computer generated faces, but did not impact the attractiveness and trustworthiness rating. Since a similar manipulation has been shown to increase the effects of social categorization in previous research (Derks et al., 2014), it is probable that the current manipulation was not strong enough. Providing different manipulation stories might nevertheless have an impact. In the stories used in this experiment, the general attitude towards computer generated faces was aimed to be influenced. A stronger manipulation could focus more specifically on the trustworthiness of computer generated faces. It is therefore recommended for future research to include manipulation stories that are extensively examined in advance.

General Discussion

The current study aimed to explore whether the belief that a face is computer generated can impact how trustworthy and attractive it appears. By means of two experiments, strong support was established for the idea that faces labeled as computer generated are perceived as less trustworthy but not as less attractive than faces labeled as natural. These findings shed a new light on the reduced trustworthiness of computer generated faces found by Balas and Pacella (2017). It is suggested that their findings were not necessarily caused by the synthetic appearance of the computer generated faces. A preexisting bias could have been enough to elicit a more negative trustworthiness rating as a consequence of social categorization.

In the study by Bernstein et al. (2007), the effect of social categorizations on face perception appeared even when the out-group was merely classified by a bogus personality test. Considering these findings, it is questionable whether the effect found in the current study was specific to computer generated faces, or could also have been achieved by using different labels. It is therefore recommended that the current study is replicated with different labels, indicating another in- and out-group distinction. If the effect is then decreased or absent, this would indicate a strong relation between social categorization and AI specifically.

It is noteworthy that no effect was found on the attractiveness rating. This was against the initial expectation that evolutionary motives would elicit a lower attractiveness rating of the faces labeled as computer generated, based on Little (2014). More research is required to examine this finding in more detail. In the current design, the focus was not specifically on investigating the influence of the belief that a face is computer generated on the process of mate selection. Hence, factors like gender, sexual orientation and relational status were not included in the analyses. To achieve a better understanding of the impact of labeling a face as computer generated on how attractive it appears, it is recommended to include these factors in the study. Furthermore, other

characteristics that are related to mate selection could be included in future research, to examine the role of evolutionary motives in the perception of computer generated faces.

The differentiation between the effect on trustworthiness and attractiveness was not unforeseen, since earlier research on human-robot interaction has focused mainly on trustworthiness rather than attractiveness. Further research is required to further explain the distinction between the effect of the labeling on the two attributes.

Acknowledging the role of social categorization within human-robot interaction can be an important step in the further development of AI. The current study emphasizes the necessity of examining this scientific field with social constructs in mind. The perceived wariness towards robots could share certain similarities to the same attitude towards humans. Elaborating on the knowledge that is already present within social sciences could therefore form an accessible approach towards the further investigation and development of AI.

Furthermore, the current findings imply that establishing a more positive bias towards algorithms is necessary for further development of this field. In the current study, however, providing a positive text prior to the task did not have a significant effect on the ratings. More research is required to examine whether manipulating bias in a different way would have more impact.

A potential approach towards improving the attitude towards algorithms is by repeated exposure. Earlier research has shown that repeated interaction with a robot can increase its likability and decrease the perceived threat and discomfort (Paetzel-Prüsmann et al., 2020). Further research could integrate this approach in the current design by using repeated sessions. A diminution of the difference in the trustworthiness ratings between the two labels over time would suggest that habituation is a key factor in overcoming the negative attitudes towards computers.

Moreover, the current findings unveil bias as an important factor that should be accounted for in research involving human-robot interaction. Hereby, the findings of Ishowo-Oloko et al. (2019) are supported. The resistance towards cooperating with a computer only when its nature is revealed, corresponds to the effect that labeling a face as computer generated makes it appear less trustworthy. In settings were a person is aware of the true nature of a robot, this awareness should be recognized as a factor influencing the findings.

In the current study, however, only explicit bias towards computer generated faces was examined. Participants were explicitly informed about the supposed nature of the faces. A more accurate representation of the effect of bias can be acquired by applying a design that allows for more implicit manipulation. By making the manipulation implicit, the task is more closely related to a realistic situation. When seeing a face in real life, no label is provided. People are often not explicitly aware of their bias. Therefore, in real life situations, the judgement of a face is influenced by implicit attitudes. Including this in the experimental design, would thus increase the ecological validity of the experiment. This could be an important starting point for future research on this topic.

Measuring the effect of implicit bias can be performed in multiple ways. Firstly, priming can be applied for this purpose. Including this in the design of the experiment, would require an expansion of the current study. After the presentation of the labeled sets of faces, a third set could be presented without any label. This set could be comprised of new faces morphed with faces from the original sets. By morphing two faces together, a new face is created with a certain similarity to each of the two faces. Participants will not be actively aware of the fact that the face looks partly familiar. However, their rating is likely to be influenced by the label that the familiar faces had in the previous presentation. Therefore, the face is implicitly paired with a label without the participant being aware of this. A similar approach has previously been used to show a

preference for political candidates that were facially similar to the participant (Bailenson et al., 2008). In this experiment, the face of the participant was morphed with the face of an unfamiliar candidate. Participants preferred this candidate over a candidate whose face was morphed with that of other voters without being consciously aware of the similarity manipulation.

Another approach to measuring implicit bias is by using the Implicit Association Test (Nosek et al., 2004). In this test, reaction time is measured to identify implicit associations. The Implicit Association Test has been used by White and White (2006) to demonstrate implicit occupational gender stereotypes. In their study, reaction times were faster when an occupation required the same response key as the gender it is stereotypically associated with. Moreover, this test has already been successfully applied in the field of robotics (Erel et al., 2019).

In the context of the current study, the test can be adapted to examine the implicit association between computer generated faces and untrustworthiness. In the first phase of this test, participants would be asked to press a certain key when a computer generated face is presented and another when a natural face is presented. Afterwards, this task will be repeated with the same two keys, but different instructions. Participants are asked to press one key when a word is presented that represents trustworthiness and the other when the word represents untrustworthiness. Subsequently, the two tasks will be combined. It is expected that response times are shorter when the same response key is shared by two concepts that are implicitly associated. Thus, a shorter response time when the computer generated faces require the same response key as the items representing untrustworthiness, would indicate an implicit bias.

The current study, therefore, indicates an interesting direction for future research. The findings provide a new insight in the challenges facing human-robot interaction. By further examining the impact of bias on perception, guidelines can be provided to improve trust in computers.

References

Bailenson, J. N., Iyengar, S., Yee, N., & Collins, N. A. (2009). *Public Opinion Quarterly, 72*(5), 938-961. https://doi.org/10.1093/poq/nfn064

Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior, 77,* 240-248. https://doi.org/10.1016/j.chb.2017.08.045

Bernstein, M. J., Young, S. G., Hugenberg, K. (2007). The cross-category effect: Mere social categorization is sufficient to elicit an own-group bias in face recognition. *Psychological science, 18*(8), 706-712. https://doi.org/10.1111%2Fj.1467-9280.2007.01964.x

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16. http://doi.org/10.5334/joc.72

Derks, B., Stedehouder, J., & Ito, T. A. (2015). Social identity modifies face perception: an ERP study of social categorization. *Social Cognitive and Affective Neuroscience, 10*(5), 672-679. https://doi.org/10.1093/scan/nsu107

Dunning, D., & Balcetis, E. (2013). Wishful seeing: How preferences shape visual perception. *Current Directions in Psychological Science, 22*(1), 33-37. https://doi.org/ 10.1177%2F0963721412463693

Erel, H., Tov, T. S., Kessler, Y., & Zuckerman, O. (2019). Robots are always social: Robotic movements are automatically interpreted as social cues. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). *Association for Computing Machinery, New York, NY, USA, Paper LBW0245, 1-6.*

https://doi.org/10.1145/3290607.3312758

Gnambs, T., Appel, M. (2019). Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. *Computers in Human Behaviour, 93,* 53-61. https://doi.org/10.1016/j.chb.2018.11.045

Ishowo-Oloko, F., Bonnefon, J., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency-efficiency tradeoff in human-machine cooperation. *Nature Machine Intelligence, 1,* 517-521. https://doi.org/10.1038/s42256-019-0113-5

Little, A. C. (2014). Facial attractiveness. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(6), 621-634. https://doi.org/10.1002/wcs.1316

Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2021). News from artificial intelligence is believed less. https://dx.doi.org/10.2139/ssrn.3787064

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47,* 1122-1135. https://doi.org/10.3758/s13428-014-0532-5

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2004). Understanding and using the implicit association test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*(2), 166-180. https://doi.org/10.1177%2F0146167204271418

Paetzel-Prüsmann, M., Perugia, G., & Castellano, G. (2020). The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*.

ACM/IEEE. https://doi.org/10.1145/ 3319502.3374786

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making, 19,* 455-468. https://doi.org/10.1002/bdm.542

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2014). Visualizing minimal ingroup and outgroup faces: Impressions, attitudes, and behavior. *Journal of Personality and Social Psychology, 106,* 897-911. https://psycnet.apa.org/doi/10.1037/a0036498

Rule, N. O., & Sutherland, S. L. (2017). Social categorization from faces: Evidence from obvious and ambiguous groups. *Current Directions in Psychological Science, 26*(3), 231-236. https://doi.org/10.1177%2F0963721417697970

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52,* 113-117. https://doi.org/10.1016/j.jesp.2014.01.005

White, M. J., & White, G. B. (2006). Implicit and explicit occupational gender stereotypes. *Sex Roles, 55,* 259-266. https://doi.org/10.1007/s11199-006-9078-z

Zebrowitz, L. A., & Montepare, J. M. (2008). Social Psychological Face Perception: Why Appearance Matters. *Social and personality psychology compass, 2*(3), 1497. https://doi.org/10.1111/j.1751-9004.2008.00109.x

Appendix A

Participants per Age Category in Experiment 1

Age category	Number of participants
< 18	1
18 -24	27
25 – 34	10
35 – 44	0
45 – 54	4
55 – 64	17
65 – 74	0
75 – 84	0
85 >	1

Appendix B

Codes of Faces per Quadrant

Code	*Attractiveness High/Low*	*Trustworthiness High/Low*	*Attractiveness rating*	*Trustworthiness rating*
BF-218	High	High	5.1	4.3
BM-043	High	High	4.9	4.2
LF-249	High	High	5.3	4.3
WF-024	High	High	4.8	3.7
WF-233	High	High	5.5	4.2
WF-242	High	High	5.8	4.2
BF-215	High	Low	3.9	2.9
BM-248	High	Low	4.4	3.1
LF-206	High	Low	4.2	2.9
LF-225	High	Low	4.0	3.1
WF-206	High	Low	4.5	2.9
WM-242	High	Low	4.0	3.1
AM-242	Low	High	2.4	4.2
BF-209	Low	High	2.4	3.8
BF-211	Low	High	2.5	3.9
BM-012	Low	High	2.5	3.8
BM-216	Low	High	2.2	3.9
LF-220	Low	High	2	3.9

BF-038	Low	Low	1.9	2.8
BF-200	Low	Low	1.6	2.4
BF-224	Low	Low	2.0	2.5
BM-213	Low	Low	2.1	2.7
LM-203	Low	Low	2.1	2.5
WF-210	Low	Low	1.9	2.7

Appendix C

Cover Story Experiment 1

Original Dutch Cover story

Voor mijn bachelor scriptie doe ik onderzoek naar de perceptie van computer gegenereerde gezichten. Dit zijn gezichten die niet aan echt bestaande mensen toebehoren, maar volledig door de computer gemaakt zijn. De technologie op dit gebied is al in een ver stadium, waardoor er extreem realistische gezichten gemaakt kunnen worden met behulp van software. Uit onderzoek is dan ook gebleken dat het erg lastig is voor mensen om een onderscheid te maken tussen computer gegenereerde en natuurlijke gezichten. Het is dan ook van belang om verder te onderzoeken op welke gebieden dit onderscheid juist wel en op welke gebieden juist niet gemaakt wordt in onze waarneming. Uit deze lijn van onderzoek is al gebleken dat op bepaalde kenmerken een groot verschil wordt gevonden in de beoordeling van computer gegenereerde en natuurlijke gezichten. Bij andere kenmerken worden daarentegen beide soorten gezichten vergelijkbaar beoordeeld. Ik wil deze bevindingen graag uitbreiden naar twee verdere kenmerken: betrouwbaarheid en aantrekkelijkheid.

Om dit te verkennen, wil ik je een aantal computer gegenereerde gezichten en een aantal natuurlijke gezichten laten zien en je vragen deze te beoordelen op deze twee kenmerken.

English translation

For my bachelor thesis, I'm investigating the perception of computer generated faces. These are faces that do not belong to real existing people, but are completely created by a computer. The technology in this field is already highly developed, due to which extremely realistic faces can be created using software. Research has shown that it is very difficult for people to distinguish

computer generated faces from natural faces. Therefore, it is of importance to further investigate on which areas this distinction can and cannot be made in our perception. This line of research has already shown that on certain characteristics a great difference is found in the rating of computer generated and natural faces. On other characteristics, however, both types of faces are rated similarly. I would like to extend these findings to two further characteristics: trustworthiness and attractiveness.

To explore this, I would like to show you some computer generated faces and some natural faces and ask you to rate these on these two characteristics.

Appendix D

Participants per Age Category in Experiment 2

Age category	Number of participants
< 18	7
18 -24	85
25 – 34	20
35 – 44	1
45 – 54	7
55 – 64	2
65 – 74	2
75 – 84	0
85 >	0

Appendix E

Manipulation Stories Experiment 2

Positive story

Original Dutch

Voor mijn bachelor scriptie doe ik onderzoek naar de perceptie van computer gegenereerde gezichten. Dit zijn gezichten die niet aan echt bestaande mensen toebehoren, maar volledig door de computer gemaakt zijn. De technologie op dit gebied is al in een ver stadium, waardoor er extreem realistische gezichten gemaakt kunnen worden met behulp van software. Uit onderzoek is dan ook gebleken dat het erg lastig is voor mensen om een onderscheid te maken tussen computer gegenereerde en natuurlijke gezichten.

Dit brengt dan ook veel mogelijkheden met zich mee. Zo kan deze technologie toegepast worden in de zorg, het onderwijs of als bestrijding van eenzaamheid. Er is al gebleken dat het gebruik van online avatars onder andere kinderen met diabetes kan helpen om te gaan met hun ziekte. Door middel van spelletjes worden deze kinderen gemotiveerd meer te leren over suikerziekte. Dit heeft een positief effect op hun zelfvertrouwen en maakt het voor kinderen makkelijker om over hun ziekte te praten. Het geven van een menselijk gezicht aan deze avatars, kan deze positieve effecten nog verder vergroten. Door deze toepassing is sociaal contact altijd binnen handbereik en kan bovendien de druk op de zorg verlicht worden.

Om te onderzoeken of computer gegenereerde gezichten hiervoor geschikt zijn, ben ik benieuwd hoe deze gezichten waargenomen worden. In het bijzonder wil ik kijken naar de betrouwbaarheid en aantrekkelijkheid. Om dit te verkennen, wil ik je een aantal computer gegenereerde gezichten

en een aantal natuurlijke gezichten laten zien en je vragen deze te beoordelen op deze twee kenmerken.

English translation:

For my bachelor thesis, I'm investigating the perception of computer generated faces. These are faces that do not belong to real existing people, but are completely created by a computer. The technology in this field is already highly developed, due to which extremely realistic faces can be created using software. Research has shown that it is very difficult for people to distinguish computer generated faces from natural faces.

This offers a lot of possibilities. Hence, this technology can be applied in healthcare, education or the fight against loneliness. It has already been shown that the use of online avatars can among other things help children with diabetes in coping with their disease. Through games, these children are motivated to learn more about diabetes. This has a positive effect on their self-confidence and makes it easier for the children to talk about their disease. Providing a human face for these avatars, can increase these positive effects even more. Due to this application, social contact is always within reach and the pressure on the healthcare sector can be relieved.

To investigate whether computer generated faces are eligible for this service, I am curious how these faces are perceived. I particularly want to look at the trustworthiness and attractiveness. To explore this, I would like to show you some computer generated faces and some natural faces and ask you to rate these on these two characteristics.

Negative story

Original Dutch:

Voor mijn bachelor scriptie doe ik onderzoek naar de perceptie van computer gegenereerde gezichten. Dit zijn gezichten die niet aan echt bestaande mensen toebehoren, maar volledig door de computer gemaakt zijn. De technologie op dit gebied is al in een ver stadium, waardoor er extreem realistische gezichten gemaakt kunnen worden met behulp van software. Uit onderzoek is dan ook gebleken dat het erg lastig is voor mensen om een onderscheid te maken tussen computer gegenereerde en natuurlijke gezichten.

Dit brengt dan ook grote risico's met zich mee. Als het onderscheid tussen wat echt is en wat nep is niet meer gemaakt kan worden, ligt misbruik op de loer. Zo kan deze technologie gebruikt worden voor chantage en vervalsing in de politiek, het bedrijfsleven en het rechtssysteem. Er is al gebleken dat er extreem realistische beelden gemaakt kunnen worden waarbij het gezicht van een bekend persoon gebruikt wordt. Een filmpje waarin een belangrijk politicus grensoverschrijdende uitspraken doet, kan enorme sociale onrust veroorzaken. Wanneer het niet te onderscheiden is of het om een vervalst filmpje gaat, zal dit alleen maar erger worden. De ontwikkeling van deze technologie kan dus leiden tot grote bedreiging.

Om te onderzoeken hoe het onderscheid gemaakt kan worden tussen natuurlijke en computer gegenereerde gezichten, ben ik benieuwd hoe deze gezichten waargenomen worden. In het bijzonder wil ik kijken naar de betrouwbaarheid en aantrekkelijkheid. Om dit te verkennen, wil ik je een aantal computer gegenereerde gezichten en een aantal natuurlijke gezichten laten zien en je vragen deze te beoordelen op deze twee kenmerken.

English translation:

For my bachelor thesis, I'm investigating the perception of computer generated faces. These are faces that do not belong to real existing people, but are completely created by a computer. The technology in this field is already highly developed, due to which extremely realistic faces can be created using software. Research has shown that it is very difficult for people to distinguish computer generated faces from natural faces.

This entails great risks. If the distinction between what is real and what is fake cannot be made anymore, misuse is waiting. Hence, this technology can be used for blackmail and fraud in politics, the corporate sector and the legal system. It has already been shown that extremely realistic images can be created in which the face of a celebrity is used. A movie in which an important politician makes provoking statements, can cause enormous social agitation. When it is not distinguishable whether the movie is forged, this will only get worse. The development of this technology can therefore result in great threat.

To investigate how natural and computer generated faces can be distinguished, I am curious how these faces are perceived. I particularly want to look at the trustworthiness and attractiveness. To explore this, I would like to show you some computer generated faces and some natural faces and ask you to rate these on these two characteristics.