Faces Merely Labelled as Artificial are Trusted Less

Baptist Liefooghe[1] *, Manuel Oliveira[1*], Luca M. Leisten[1,2], Eline Hoogers[1], Henk Aarts[1], Ruud Hortensius[1]

[1] Department of Psychology, Utrecht University

[2] Radboud University

Author note

Part of this research was supported by Human-AI Alliance at Utrecht University. Baptist Liefooghe and Manuel Oliveira contributed equally to this work.

* Correspondence concerning this article should be addressed to Baptist Liefooghe, b.liefooghe@uu.nl, or Manuel Oliveira, m.j.barbosadeoliveira@uu.nl

Running head: Trust in artificial faces

Competing interests

The authors declare no competing interests.

Data accessibility statement

Data and materials can be found on this paper's project page on the Open Science Framework:

https://osf.io/3hrx2/?view_only=7e4a7880e823439fab27177ecbb56664

Abstract

Artificial intelligence plays a crucial role on our daily lives. At the same time, artificial intelligence is often met with reluctance and distrust. Previous research demonstrated that faces that are visibly artificial are considered to be less trustworthy and remembered less accurately compared to natural faces. Current technology, however, enables the generation of artificial faces that are indistinguishable from natural faces. Accordingly, we tested whether natural faces that are merely labelled to be artificial are also trusted less. In three experiments (*N* = 399), we observed that natural faces merely labeled as being artificial were judged to be less trustworthy. This bias was robust and did not depend on the degree of trustworthiness and attractiveness of the faces, nor could it be modulated by changing raters' attitude towards artificial intelligence. At the same time, we did not observe differences in recall performance. We conclude that understanding and changing social evaluations towards artificial intelligence goes beyond eliminating physical differences between artificial and natural entities.

*Keywords***:** artificial intelligence, trust, face perception, outgroup effects, social psychology

Faces Merely Labelled as Artificial are Trusted Less

Artificial intelligence (AI) applications support many processes in today's society, such as entertainment, service industry, administration, governance, transportation and health care (Abduljabbar et al., 2019; Hamet & Tremblay, 2017; Huang & Rust, 2018; Wirtz et al., 2018). However, AI solutions are often met with reluctance by target users, which jeopardizes the applicability of these systems (Davenport & Ronanki, 2018). A key determinant of Human-AI interaction is trust. Trust predicts the use of AI and an optimal level of trust is essential since low trust can lead to bias and disuse, and over-trust can lead to misuse of AI (Lee & See, 2004; Parasuraman & Manzey, 2010). On the one hand, trust in AI depends on relatively objective features of an AI application such as its performance (such as reliability, error rate, dependability), automation and transparency (for reviews, see Glikson & Woolley, 2020; Hancock et al., 2011; Hoff & Bashir, 2015), or the task user and system are involved in (e.g., task difficulty or workload, Hoff & Bashir, 2015). On the other hand, trust in AI is also driven by expectations and beliefs the user holds about AI. An interesting finding demonstrating the importance of user' attitudes in Human-AI interaction is that synthetic or computer-generated faces are judged to be less trustworthy compared to natural faces (Balas & Pacella, 2015, 2017). This difference indicates that a bias against AI thus exists even at early stages of impression formation. Here, we show this bias pertains even for artificial faces that are undistinguishable from real faces.

Trustworthiness judgments on the basis of faces follow from our expertise in inferring traits from facial features, which in turn depends on the degree to which we are exposed to a particular face (e.g., Dotsch et al., 2016; Ng & Lindsay, 1994). Previous research suggests that such face expertise is less developed for computer-generated faces (e.g., Crookes et al., 2015), because these faces are less common in our daily environment. Artificial faces are also less well-remembered compared to human faces (see also Balas & Pacella, 2015; Crookes et al., 2015). These findings suggest that – from the perspective of a human observer – artificial faces

constitute an outgroup compared to real human faces, which can lead to a form of the other-ethnicity effect (see Meissner & Brigham, 2001 for a review) that impacts how well these faces are remembered (Balas & Pacella, 2015), discriminated (Balas & Tonsager, 2014), or socially evaluated (Birkás et al., 2014; Stanley et al., 2012). As a result, the difference in appearance between natural and computer-generated faces may affect the extent to which computer-generated faces are considered to be trustworthy compared to natural faces.

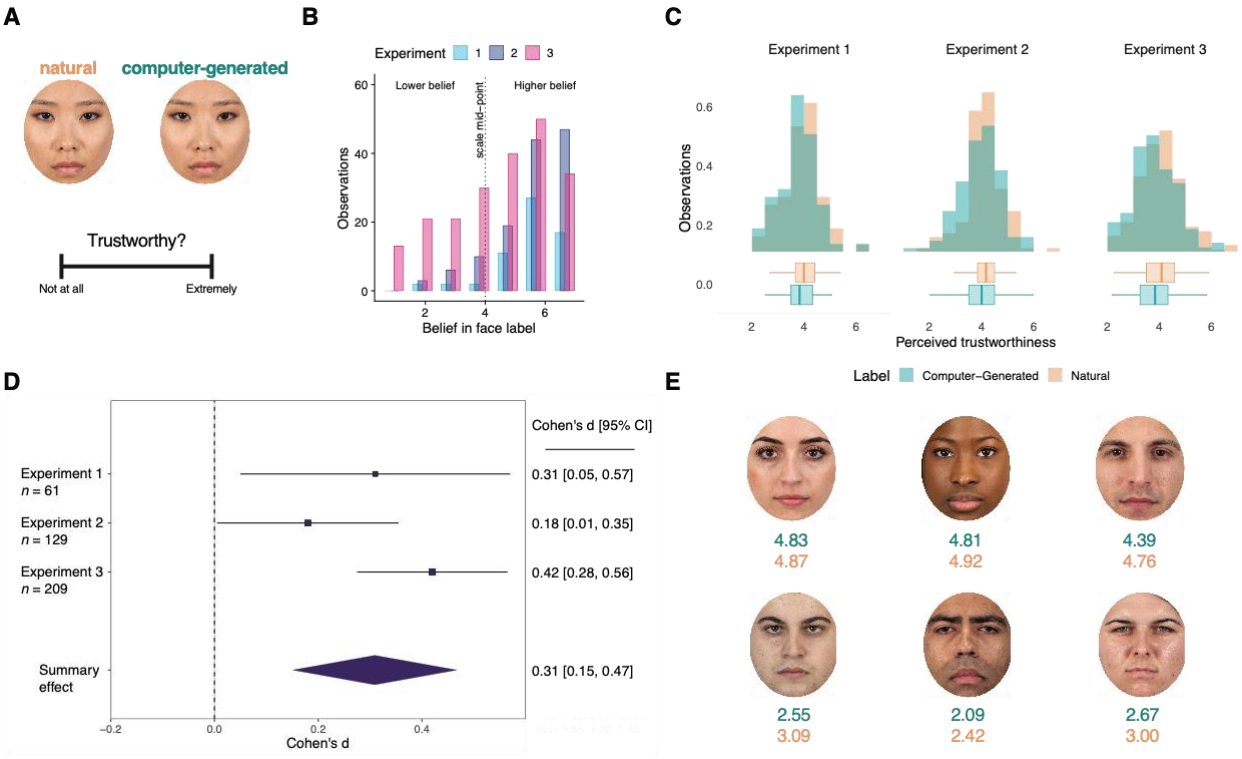
The difference in trustworthiness between natural and computer-generated faces bear important implications when considering AI applications that use artificial faces to interact with humans, such as in therapeutic or educational settings (Billard et al., 2007; Matarić et al., 2009; Paiva et al., 2004). At the same time, the question arises whether such distrust in artificial faces is limited to situations in which these faces appear to be synthetic or if this distrust is also present when artificial faces are undistinguishable from real human faces. On the one hand, current technology enables to render faces that look completely realistic (see for instance, https://thispersondoesnotexist.com/). On the other hand, outgroup effects are not only due to differences in face expertise, but are also related to differences in social cognitions typically elicited when processing in- and outgroup members (see Sporer, 2001, for a review). Merely categorizing a stimulus as an ingroup or an outgroup member impacts how this stimulus is subsequently processed (Tajfel, 1982; Tajfel et al., 1971; Tajfel & Turner, 1986). For instance, Bernstein, Young, and Hugenberg (2007) presented students a series of faces presented on red or green backgrounds. Participants were instructed that faces on the red background were university ingroup members and that faces on the green background were university outgroup members. Despite students and faces were from the same ethnic group, thus controlling for face expertise, ingroup faces were remember more accurately than outgroup faces. Such social-categorization effect may reflect differences in processing mode between in- and outgroup faces, with less attention being paid to the discriminative features of outgroup faces in comparison to ingroup faces (e.g., Levin, 1996; MacLin & Malpass, 2001, 2003). Alternatively, it

may be a motivational effect with less effort made to encode outgroup compared to ingroup faces (e.g., Rodin, 1987).

The above considerations indicate that differences in face expertise and differences in physical appearance between faces are not a prerequisite to observe outgroup effects. Based on the observation that an outgroup can be created on the mere basis of (arbitrary) social categorization (Bernstein et al., 2007) and based on the hypothesis that artificial faces form an outgroup that is trusted less compared to natural faces (Balas & Pacella, 2017), we hypothesized that faces labelled to be artificial will be judged to be less trustworthy compared to faces labelled to be real, even when these faces are undistinguishable. In order to test this hypothesis, the present study used a research approach similar to Bernstein et al. (2007). Participants were always presented with real natural faces, some of which were randomly labelled as being computer-generated and others were randomly labelled as being natural. We predicted that faces merely labelled as being computer-generated would be rated as less trustworthy, even though all faces presented were actually real. We present three experiments in which this hypothesis is confirmed (see Figure 1).

Figure 1

Overview of Results for the Three Experiments.



Note. (A) Example of the face rating task; (B) Results for believability in the face label manipulation. (C) Distribution of trustworthiness ratings by face label for each experiment. (D) Average effect size of the “mere-label” effect within and across experiments. (E) Face exemplars with respective average trustworthiness ratings by label.

Experiment 1

In order to test our hypothesis that merely labelling a face as being computer-generated is sufficient for that face to be considered as less trustworthy, we selected natural faces from the Chicago Face Database (Ma et al., 2015) and either labelled them as being natural or computer-generated. Participants judged the faces' trustworthiness with 7-point Likert scale. In addition, facial judgements of attractiveness were assessed. Attractiveness judgements are mainly based on a global affective response that requires minimal inferential activity (e.g., Zajonc, 1980) and offers a benchmark for more sophisticated judgments such as of trustworthiness (see also Willis & Todorov, 2006). Finally, we also controlled whether potential differences in judgments between faces labelled as computer-generated or natural depends on the degree of attractiveness or trustworthiness associated with that face. To this end, we selected high and low trustworthy faces as well as high and low attractive faces. This resulted in four groups of stimuli defined by crossing the extreme poles of facial trustworthiness and facial attractiveness, namely: trustworthy and attractive, untrustworthy and unattractive, trustworthy and unattractive, untrustworthy and attractive.

Method

Participants. A sample of 60 participants was recruited (48 female; age categories: <18yrs: n = 1; 18-24yrs: n=27; 25-34yrs: n=10; 45-54yrs: 4; 55-64yrs: 17; and >85yrs: n=1). Participants belonged to the social network of one of the authors and participated for free. To obtain a power of .80 for a medium-sized effect (d= 0.5) in a within-subjects design, a minimum sample of 32 participants was needed. The sample size is thus sufficiently large for our research purpose.

Materials. A survey was created using Qualtrics online software (https://www.qualtrics.com). This survey consisted of 24 pictures of natural faces from the Chicago Face Database (Ma et al., 2015), a free resource consisting of 158 high-resolution, standardized photographs of Black and White males and females between the ages of 18 and 40 years. Based on the trustworthiness and attractiveness ratings provided in this database (using a 7-point Likert scale), we selected faces with the most extreme scores on both dimensions such that we obtained four categories of 6 faces each: high trustworthy (*M*= 5.23; *SD*= 0.38) and high attractive faces (*M*= 4.15; *SD*= 0.23); high trustworthy (*M*= 4.17; *SD*= 0.24) and low attractive faces (*M*= 3.00; *SD*= 0.11); low trustworthy (*M*= 2.25; *SD*= 0.26) and high attractive (*M*= 3.68; *SD*= 0.44); low trustworthy (*M*= 1.93; *SD*= 0.19) and low attractive (*M*= 2.60; *SD*= 0.15). Reference codes and ratings of the selected faces are presented in https://osf.io/3hrx2/?view_only=7e4a7880e823439fab27177ecbb56664. In order to ensure that participants rated the faces proper, they were cropped in oval shape, excluding hair and clothing. In each category, faces were randomly divided in two sets of 3 faces. These sets were either labelled as being natural or computer-generated faces. This labelling was counterbalanced over participants.

Procedure. The survey started with a cover story in which it was emphasized that current computer capabilities permit to render faces that are almost undistinguishable from natural faces and that research is needed to investigate characteristics on which differences between natural and computer-generated faces can be distinguished. An English translation of this cover story is provided at https://osf.io/3hrx2/?view_only=7e4a7880e823439fab27177ecbb56664. Following the instructions, informed consent, age category, and gender identification were asked.

Participants were either first presented with the set of faces labelled as natural or with the set of faces labeled as computer-generated. Each set consisted of 12 faces (3 faces per

category). Within each set faces were presented in a random order, one at a time in the middle of the screen. Below each face, two 7-point Likert scales were presented. One for attractiveness and one for trustworthiness. For both scales the right side was labeled with “absolutely not” (1) and the right side with “extremely” (7). The up-down order of both scales varied per face. Following the rating of the first set of faces, the second set was presented. The presentation order of both sets of faces (e.g., natural faces first, computer-generated faces second) was counterbalanced over participants. Participants were also informed about the nature of the upcoming set of faces (i.e., natural or computer-generated). After rating both sets of faces, participants were asked to indicate how strongly they believed that the computer-generated faces were actually generated by a computer, using a 7-point scale ranging from 1 (Absolutely Not) to 7 (Extremely).

Attractiveness and trustworthiness ratings were subjected to a 2 (Facial Attractiveness: High vs. Low) by 2 (Facial Trustworthiness: High vs. Low) by 2 (Label: Natural, Computer-generated) repeated measures ANOVA. All data processing and analyses were performed by using R (R Core Team, 2017). ANOVAs were calculated by using “afex” (Singmann et al., 2021) and follow-up contrasts on the model estimates by using “phia” (De Rosario-Martinez, 2015). Raw data and corresponding analysis scripts are available at https://osf.io/3hrx2/?view_only=7e4a7880e823439fab27177ecbb56664.

Results

Participants’ overall belief that the computed-generated faces were actually generated by a computer was 5.78 (*SD*= 1.18). This was significantly higher than the middle of the Likert scale, $t(60)= 11.93$, $p< .001$.

Trustworthiness. Faces labelled to be computer-generated (*M*= 3.82; *SE*= .09) were rated to be less trustworthy compared to faces labelled to be natural (*M*= 3.97; *SE*= .09), $F(1, 60)= 5.78$, *MSe*= .48, $p < .05$, $\eta_p^2= .09$. High trustworthy faces (*M*= 4.36; *SE*= .08) were rated to

be more trustworthy than low trustworthy faces (M= 3.43; SE= .10), F(1, 60)= 192.38, MSe= .54, p < .001, η_p^2= .76. High attractive faces (M= 4.18; SE= .09) were rated to be more trustworthy compared to low attractive faces (M= 3.61; SE= .10), F(1, 60)= 66.34, MSe= 0.59, p < .001, η_p^2= .53.

Nor the interaction between Label and Facial Trustworthiness, neither the interaction between Label and Facial Attractiveness, were significant, both Fs < 1. Similarly, the interaction between Facial Attractiveness and Facial Trustworthiness, F(1, 60)= 2.99, MSe= .28, p = .09, η_p^2= .05, as well as the three-way interaction were not significant.

Attractiveness. Mean attractiveness ratings did not differ reliably between both labels (natural: M= 2.99; SE= .09; computer-generated: M= 3.07; SE= .09), F(1, 60)= 1.44, MSe= .50, p = .24, η_p^2= .02. High attractive faces (M= 3.93; SE= .11) were rated to be more attractive compared to low attractive faces (M= 2.13; SE= .09), F(1, 60)= 349.12, MSe= 1.15, p < .001, η_p^2= .85. High trustworthy faces (M= 3.44; SE= .10) were also rated to be more attractive than low trustworthy faces (M= 2.63; SE= .09), F(1, 60)= 175.10, MSe= .46, p < .001, η_p^2= .75.

Facial Attractiveness and Facial Trustworthiness interacted, F(1, 60)= 25.29, MSe= .33, p < .001, η_p^2= .30. The effect of facial trustworthiness on the mean ratings was more pronounced for the high attractive than for low attractive faces. None of the remaining interactions was significant, all Fs < 1.

Discussion

The results of Experiment 1 suggest that faces labelled as computer-generated are judged to be less trustworthy than faces labelled as natural. This difference did not depend on the degree of attractiveness or trustworthiness associated with that face. For the attractiveness ratings we did not observe reliable differences between so-called computer-generated and natural faces. Experiment 1 offers first evidence that merely instructing that faces are computer-generated makes them less trustworthy compared to faces told to be real. This finding

extends previous results by indicating that differences in physical appearance between computer-generated and natural faces are not a prerequisite to obtain a difference in perceived trustworthiness.

Experiment 2

Experiment 2 aims to replicate and extend the findings of Experiment 1 by testing the robustness of the bias against computer-generated faces. Biases against well-established social outgroups, such as in the case of racial prejudice, are difficult to reduce through interventions (see Jackson, 2018; Lai et al., 2016; Van Dessel et al., 2020 for a discussion). In contrast, attitudes to unfamiliar and new social categories, such as members of fictious tribes, are easily malleable, for instance by using instructions (e.g., De Houwer, 2006; Gregg et al., 2006; Van Dessel et al., 2015, 2020). Based on these considerations Experiment 2 tested whether the bias against computer-generated faces could be modulated by manipulating the nature of a cover story presented at the start of the experiment. Two cover stories were used, which differed in spin. The positive story emphasized the benefits of using realistic computer-generated faces that could not be distinguished from natural faces (e.g., the use of virtual assistants in clinical and educational settings), whereas the negative story emphasized the treat of realistic computer-generated faces (e.g., deep fakes). We tested whether these cover stories were sufficient to modulate the bias towards computer-generated faces.

Method

A sample of 130 participants was recruited through social media (100 female, 29 male; age categories: <18yrs: n = 8; 18-24yrs: n= 86; 25-34yrs: n= 23; 35-44yrs: n= 1; 45-54yrs: n= 8; 55-64yrs: n= 2; and 65-74yrs: n= 2). A gift voucher of 25 euro was allotted to motivate participation. One participant had a missing value and was excluded from data analysis. Sixty-three participants received the positive story, and 66 participants received the negative story.

The sample size in each condition was largely sufficient to detect a medium-sized effect (d= .5) with a power of .80.

Stimuli, materials and procedure were similar to Experiment 1. English translations of the two cover stories are available at https://osf.io/3hrx2/?view_only=7e4a7880e823439fab27177ecbb56664. Following the rating of the faces, the extent to which participants believed that the computer-generated faces were actually generated by a computer was assessed. In addition, participants also rated their attitude towards artificial intelligence in general on a 7-point Likert scale. This additional question aimed to measure whether different attitudes towards artificial intelligence were induced by both cover stories.

Data analysis. Attractiveness and trustworthiness ratings were subjected to a 2 (Story: Positive, Negative) by 2 (Facial Attractiveness: High, Low) by 2 (Facial Trustworthiness: High, Low) by 2 (Label: Natural, Computer-generated) mixed ANOVA with repeated measures on the last three factors.

Results

Participants' overall belief that the computer-generated faces were actually generated by a computer was 5.83 (*SD*= 1.26), which was significantly higher than the middle of the Likert scale, $t(128)= 16.45$, $p < .001$. In addition, participants were significantly more positive towards artificial intelligence in the condition in which a positive cover story was used (*M*= 4.27; *SE*= .16) compared to the condition in which a negative cover story was used (*M*= 3.80; *SE*= .15), $F(1, 127)= 4.53$, *MSe*= 1.55, $p < .05$, $\eta_p^2 = .03$.

Trustworthiness. Faces labelled to be computed-generated (*M*= 3.98; *SE*= .07) were rated to be less trustworthy compared to faces labelled to be natural (*M*= 4.11; *SE*= .06), $F(1, 127)= 4.22$, *MSe*= 1.15, $p < .05$, $\eta_p^2 = .04$. There was no main effect of Cover Story, $F<1$

(Positive: *M*= 4.03, *SE*= .08; Negative: *M*= 4.02, *SE*= .08). Cover Story did not interact with Label, *F* <1.

High trustworthy faces (*M*= 4.51; *SE*= .06) were rated to be more trustworthy than low trustworthy faces (*M*= 3.58; *SE*= .06), *F*(1, 127)= 328.40, *MSe*= .67, $p < .001$, $\eta_p^2 = .72$. High attractive faces (*M*= 4.23; *SE*= .06) were rated to be more trustworthy compared to low attractive faces (*M*= 3.86; *SE*= .07), *F*(1, 127)= 69.85, *MSe*= 0.51, $p < .001$, $\eta_p^2 = .36$. None of the interactions was significant. The largest F-value was observed for the interaction between Facial Attractiveness and Facial Trustworthiness, *F*(1, 127)= 2.22, *MSe*= 0.39, $p = .14$, $\eta_p^2 = .02$.

Attractiveness. Mean attractiveness rating did not differ reliably between faces labelled as natural (*M*= 3.31; *SE*= .07) and faces labelled as computer-generated (*M*= 3.24; *SE*= .07), *F*(1, 127)= 2.75, *MSe*= .50, $p = .10$, $\eta_p^2 = .02$.

Mean ratings did not differ significantly as a function of Story, *F*<1 (Negative: *M*= 3.32, *SE*= .08; Positive: *M*= 3.28, *SE*= .08). Story did not interact significantly with Label, *F*(1, 127)= 1.09, *MSe*= 1.15, $p = .30$, $\eta_p^2 = .01$. Other interactions involving the factor Label were also not significant, all *Fs* < 1.

High attractive faces (*M*= 4.11; *SE*= .08) were rated to be more attractive compared to low attractive faces (*M*= 2.45; *SE*= .07), *F*(1, 127)= 538.30, *MSe*= 1.32, $p < .001$, $\eta_p^2 = .81$. High trustworthy faces (*M*= 3.61; *SE*= .07) were also rated to be more attractive than low trustworthy faces (*M*= 2.94; *SE*= .07), *F*(1, 127)= 320.20, *MSe*= .36, $p < .001$, $\eta_p^2 = .72$. Facial Attractiveness and Facial Trustworthiness interacted, *F*(1, 127)= 32.78, *MSe*= .28, $p < .001$, $\eta_p^2 = .21$. The difference between high and low trustworthy faces was less pronounced for the low attractive faces compared to the high attractive faces.

Discussion

Similar to Experiment 1, faces labelled as computer-generated were rated as being less trustworthy compared to faces labelled as natural. This bias did not depend on the degree of

attractiveness or trustworthiness associated with that face and did not differ between cover stories. Yet, these stories had a significant effect on the attitude of participants towards AI. The negative story led to less positive attitudes compared to the positively themed story. Our cover stories thus impacted participants attitudes towards AI, but not the difference in perceived trustworthiness between computer-generated and natural labelled faces. For the attractiveness ratings no reliable difference was observed between both types of faces.

Taken together, Experiment 2 replicates Experiment 1 and again demonstrates that faces believed to be computer-generated are considered less trustworthy and this even if these faces are physically undistinguishable from natural faces (cf. Balas & Pacella, 2017). In line with biases against well-established social outgroups (Jackson, 2018; Lai et al., 2016; Van Dessel et al., 2020 for a discussion), this difference in trustworthiness could not be modulated by simply adding contextual information.

Experiment 3

Experiment 3 aims to replicate the findings of Experiment 2 with sufficient statistical power to detect a potentially smaller (i.e., $d = 0.4$) moderation of the bias towards computer-generated faces by background story. The second aim of Experiment 3 was to test if the memory effect associated with the recognition of outgroup faces that is typically found in the literature also emerges for faces labeled as artificial. In a face memory task, faces categorized as belonging to an outgroup (e.g., faces from another ethnicity than the perceiver's) tend to receive a lower proportion of hits and elicit a higher proportion of false alarms (see Meissner et al., 2005; Meissner & Brigham, 2001). In other words, outgroup faces tend to be less well identified as having been encountered before (i.e., during the study phase) and are less well differentiated from faces that were not presented before. An explanation for the increased proportion of false alarms is based on the idea that the encoding of faces in memory is optimized to facilitate the discrimination between faces to which we are more frequently

exposed to (Valentine, 1991; Valentine et al., 2016). However, findings have been mixed in studies testing the hypothesis that artificial faces are harder to remember compared to natural faces. Some studies encounter a superior performance to remember natural faces compared to their computer-generated versions (Balas & Pacella, 2015; Crookes et al., 2015), but do not find a higher tendency to commit false alarms for (outgroup) computer-generated faces. Yet, other studies encounter no difference in the ability to remember computer-generated or natural faces, and find instead a higher tendency for computer-generated faces to elicit false alarms (Kätsyri, 2018). In addition, Bernstein et al. (2007) observed impeded recall performance for faces arbitrarily labelled as belonging to an outgroup. Based on these previous studies and on the current observation that simply labelling a face to be computer-generated results in lower trustworthiness ratings, we tested whether natural faces labelled to be computer-generated are also less well remembered. Such a result would suggest that our labelling manipulation also leads to an outgroup bias.

Method

This experiment was pre-registered (protocol: https://osf.io/w4bca). An initial sample of 210 participants was recruited via Prolific Academic (www.prolific.co). Participants were fluent in English and resided in 21 different countries. One participant was excluded from the analysis due to zero variability in the ratings The final sample included in the analyses consisted of 209 participants (104 female, 105 male; median age (IQR): 24 (8), age range: 18-64). The pre-registered sample size of 208 was overshot by one participant due to technicalities of the Prolific platform (i.e., automatic replacement of participant despite completion of experiment). All participants, including the excluded and additional ones, were compensated with £1.50 according to an hourly rate of £7.50.

Stimuli, materials and procedure were similar to Experiment 2's, with some exceptions. Participants were only allowed to move past the background story screen after 30 seconds had

elapsed (this control was not pre-registered and was implemented after the initial pre-registered technical check at n = 5 during data collection). The previously used 24 pictures (6 per face category; see Experiment 1) were complemented with 18 faces from the same database per category, adding up to a total of 96 faces. In the task, participants were shown a total of 24 faces: 6 per category, randomly drawn from the full set of 96 faces. These faces were randomly labeled as “computer-generated” or “natural” (counterbalanced). The trustworthiness scale was always presented under the attractiveness scale. After rating all the faces in trustworthiness and attractiveness, participants completed a surprise memory task. In this task they were shown 32 unlabeled faces: half presented before (balanced per label) and half not presented before, all balanced per category. For each face, participants indicated if they had seen the face in the previous block by clicking “yes” or “no”. Finally, as in Experiment 2, two items assessed a participant’s belief in the stimulus’ nature and their general attitude towards artificial intelligence.

For the memory task, we calculated a d-prime (d’) or sensitivity index for each participant. Sensitivity (d’) reflects a participant’s ability to discriminate between faces that were previously presented during the initial study from faces that were not. In addition, we calculated the response criterion (“c”) index. This index captures any bias towards responding that faces were presented before (less conservative criterion) or towards responding that faces were not presented before (more conservative criterion). To compute d’ and c we decomposed the accuracy of responses (1 = item is correctly recognized; 0 = item is incorrectly recognized) into hits (i.e., when a face that was previously presented, commonly designated as ‘test face’, is correctly recognized as such); misses (i.e., when a face that was previously presented is not recognized); false alarms (i.e., when a new face that was not previously presented, commonly designated as ‘lure face’, is recognized as having been presented before); and correct rejections (i.e., when a lure face is correctly identified as not having been previously presented).

To deal with the occasional cases of perfect accuracy, which lead to infinite values of d’, we employed the log-linear correction method to the hit and false alarm rates described in

Stanislaw & Todorov (1999). This correction is implemented by adding 0.5 to both the number of hits and the number of false alarms, and by adding 1 to both the number of signal trials (test faces) and the number of noise trials (lure faces) before calculating the hit and false alarm rates. We then conducted a Story by Label mixed ANOVA for d' and c.

Results

Participants' overall belief that the computer-generated faces were actually generated by a computer (M = 4.67, SE = 0.12) was again significantly higher than the middle of the scale, $t(208)$= 5.38, $p < .001$. And again, participants were significantly more positive towards artificial intelligence in the condition in which a positive cover story was used (M = 4.82, SE = 0.13) compared to the condition in which a negative cover story was used (M = 3.77; SE = 0.13), $F(1,207) = 32.14$, $MSe = 1.78$, $p < .001$, $\eta_p^2 = .134$.

Trustworthiness. The interaction between Story and Label was not significant, $F < 1$, $\eta_p^2 < .001$. Neither was the main effect of Story, $F < 1$, $\eta_p^2 = .003$. Importantly, the main effect of Label was again replicated, $F(1, 207) = 37.20$, $Mse = 0.71$, $p < .001$, $\eta_p^2 = .152$, indicating that faces labeled as computer-generated were rated as less trustworthy (M = 3.88, SE = 0.06) than faces labeled as natural (M = 4.13, SE = 0.06). Finally, trustworthy faces were rated as more trustworthy (M = 4.40, SE = 0.06) than untrustworthy faces (M = 3.60, SE = 0.06), $F(1, 207) = 294.68$, $Mse = 0.89$, $p < .001$, $\eta_p^2 = .587$, and attractive faces were rated as more trustworthy (M = 4.22, SE = 0.05) than unattractive faces (M = 3.78, SE = 0.06), $F(1, 207) = 134.92$, $Mse = 0.59$, $p < .001$, $\eta_p^2 = .395$.

Attractiveness. Contrary to the previous experiments, the pattern of results was now similar to that obtained with trustworthiness ratings. There was no interaction between Story and Label nor a main effect of Story, both Fs < 1, $\eta_p^2 < .001$. The main effect of Label was significant, $F(1, 207) = 11.06$, $Mse = 0.62$, $p = .001$, $\eta_p^2 = .05$, indicating that faces labeled as computer-generated were rated as less attractive (M = 3.60, SE = 0.06) than their natural

counterparts (M = 3.72, SE = 0.06). An additional test (non-preregistered) of the interaction between label and judgment dimension (i.e., attractiveness, trustworthiness) indicated that the difference between faces labelled as natural and faces labelled as computer-generated was larger for the trustworthiness ratings, M_{diff} = 0.25, SE = 0.41, p < .001, d = 0.85, 95% CI [0.56, 1.13], compared to the attractiveness ratings, M_{diff} = 0.13, SE = 0.38, p = .001, d = 0.46, 95% CI [0.19, 0.74], F(1, 207) = 10.33, p = .002, η_p^2 = .048.

Finally, attractive faces were rated as more attractive (M =4.28, SE = 0.06) than unattractive faces (M = 3.04, SE = 0.06), F(1, 207) = 943.23, Mse= 0.68, p < .001, η_p^2= .820, and trustworthy faces were rated as more attractive (M = 3.89, SE = 0.06) than untrustworthy faces (M = 3.43, SE = 0.06), F(1, 207) = 133.10, Mse= 0.69, p < .001, η_p^2= .391. Facial Attractiveness and Facial Trustworthiness interacted, F(1, 207)= 9.04, MSe= .38, p = .003, η_p^2= .042. The difference in rated attractiveness between high and low trustworthy faces was less pronounced for the high attractive faces compared to the low attractive faces. This pattern mirrors the ones found for the same interaction in Experiments 1 and 2, where the difference was less pronounced for low attractive faces. This could be a result of the intended variability in the content of the stimulus set, as this time the set was composed of faces randomly drawn from a larger pool of faces.

Memory task. Regarding sensitivity (d'), participants showed good ability to detect faces that had been previously shown (one-sided t-tests against zero: mean $d'_{computer\text{-}generated}$ = 0.97, SE = 0.04, t(208) = 25.49, p < .001; mean $d'_{natural}$ = 0.94, SE = 0.04, p < .001, t(208) = 25.86, p < .001). The ANOVA revealed no interaction between Story and Label, F(1, 207) = 1.81, MSe= 0.19, p = .179, η_p^2= .009. There was also no main effect of Story, F < 1, η_p^2= .004, nor Label, F < 1, η_p^2= .001. The response bias (c) analysis revealed an equal tendency to commit false alarms for faces previously labeled as computer-generated (c = -0.48, SE = 0.02) or as natural (c = -0.47, SE = 0.02), as suggested by the non-significant main effect of Label, F < 1, η_p^2= .001.

The main effect of Story, $F < 1$, $\eta_p^2 = .004$, and its interaction with Label, $F < 1$, $\eta_p^2 = .009$, were also non-significant.

Discussion

As in the previous two experiments, participants in Experiment 3 rated faces labeled as computer-generated as less trustworthy than those labeled as natural. In contrast, to the previous experiments, a similar but smaller bias was found for the attractiveness ratings. This could be expected in light of the strong positive relationship between facial judgments of trustworthiness and attractiveness (e.g., Oosterhof & Todorov, 2008; Ramos et al., 2016). As in Experiment 2, the background story did not moderate the judgement biases despite of higher control in the degree of exposure to these stories and sufficient statistical power to detect a smaller effect. The cover story only exerted an effect on the reported attitude towards AI.

The memory task indicated that faces were remembered equally well regardless of the label they were paired with at initial exposure. This suggests no differences in how the label-face pairings were encoded in memory at initial exposure. Such a finding diverges from the memory advantage for natural faces over visibly synthetic faces encountered by Balas and Pacella (2015; see also Crookes et al., 2015). Moreover, we failed to replicate the stronger tendency to commit false alarms for computer-generated stimuli previously encountered by Kätsyri (2018).

General Discussion

Previous research demonstrated that computer-generated faces are processed and judged differently than natural faces (e.g., Balas & Pacella, 2015; Crookes et al., 2015). Balas and Pacella (2017) hypothesized that differences in face expertise between computer-generated and natural faces may lead to an outgroup bias (Meissner & Brigham, 2001), resulting in less trust in computer-generated compared to natural faces. In line with their

hypothesis, Balas and Pacella (2017) observed that computer-generated faces were judged to be less trustworthy compared to natural faces. Here, we replicate and elaborate this finding in three experiments by demonstrating that merely instructing participants that a natural face is computer-generated leads participants to rate this face as being less trustworthy. This effect did not depend on the degree to which the faces were considered to be trustworthy or attractive (Experiments 1-3), is not modulated when changing attitudes towards AI by means of a cover story (Experiments 2-3) and was not associated with a memory bias (Experiment 3).

The prime conclusion present study is that a bias against computer-generated faces, such as lower trustworthiness, can be exclusively triggered by higher level social cognitive processes and does not necessarily require an explanation based on low-level perceptual mechanisms. However, we do not exclude that the synthetic appearance of computer-generated faces may contribute to the trustworthiness bias we observed and the judgement of a face follows from an interaction between top-down and bottom-up processing streams of information (Freeman & Ambady, 2011; Hehman et al., 2017).

The working hypothesis of the present study was that faces labelled to be computer-generated would be considered as an outgroup, which would render them less trustworthy. However, an important behavioral marker for outgroup effects is that outgroup faces are remembered less well. Here, we did not encounter any evidence supporting the thesis that faces categorized as computer-generated would be less well remembered than faces categorized as natural. On the one hand, such a finding does not align with previous studies that encountered a memory advantage for natural faces compared to computer-generated faces (Balas & Pacella, 2015; Crookes et al., 2015). The main difference between these studies and ours lies in the true nature of the faces being compared. While previous studies contrasted natural and visibly computer-generated faces, we only used natural faces. Possibly, memory biases for computer-generated faces only emerges when notable differences are present. Memory effects would thus be driven more by stimulus characteristics (i.e., from a bottom-up

stream of information) than by processes of social categorization (i.e., from a top-down stream of information). At the same time, it is worth noticing that some authors also failed to find a memory advantage for natural faces despite using actual computer-generated faces (Kätsyri, 2018). On the other hand, Bernstein and colleagues (2007) did observe impaired memory for faces arbitrarily categorized as an outgroup, while controlling for differences in perceptual face expertise. Although we also used arbitrary social categorization, our findings are not in line with that study neither. The absence of an outgroup effect in the present study is furthermore supported by the finding that the difference between high-low trustworthy faces and high-low attractive faces was similar for faces labelled to be natural and faces labelled to be computer-generated. It has been hypothesized that members of an outgroup are considered to be more homogenous (Park & Rothbart, 1982; Quattrone & Jones, 1980). As such, some attenuation could have been predicted, with differences in trustworthiness and attractiveness being smaller for faces labelled to be computer-generated compared to faces labelled to be natural.

Taken together, our results are difficult to reconcile with the idea of an outgroup bias. Although we remain cautious in making conclusions on the basis of null effects, an alternative explanation could be that our effect is not mediated by an outgroup bias but reflects a more general evaluative conditioning effect. Evaluative conditioning leads to a change in valence of a stimulus due to the pairing of that stimulus with another stimulus that is intrinsically negative or positive (see Hofmann et al., 2010 for a review). In the context of the current study, it is possible that labels referring to AI (e.g., computer-generated, artificial, synthetic) are sensed to be less positive compared to labels referring to real entities (e.g., human, natural, real). As such these labels may function as unconditioned stimuli, which bias attitudes towards the faces they are paired with. As a result, faces paired with labels referring to AI are perceived as being less positive and, more specially, less trustworthy. Although such account is speculative at this stage it indicates that future research is needed to further specify the processes underlying the difference in processing artificial and real faces and test the boundary conditions of these

differences. For instance, recent studies reported that state-of-the-art synthetic faces that are undistinguishable from real faces elicit higher trustworthiness ratings (Nightingale & Farid, 2022).This bias seems to depend on the degree to which these faces are believed to be real (Tucciarelli et al., 2022). Such finding corroborates with the current results by indicating the importance of beliefs and attitudes, which were explicitly manipulated in the current study by using labels.

Interestingly, with the exception of Experiment 3, we did not observe reliable differences between faces labelled as computer-generated or as natural for attractiveness ratings. Although attractiveness may be processed differently between real and synthetic faces (Balas, Tupa, & Pacella, 2018), evidence supporting clear differences in attractiveness between in- and outgroup faces is generally mixed (Burke et al., 2013; Cunningham et al., 1995; Jones, 1995; Rhodes et al., 2001, 2005). Taken into consideration that the perception of attractiveness has partly a biological basis with specialized perceptual processing that is automatic and stimulus-driven (e.g., Langlois et al., 1987; Little et al., 2011; Salvia et al., 1975), the possibility arises that our manipulation was too high-end (i.e., simply instructing social categories) to obtain reliable differences.

To conclude, the present study offers an important extension to the research on and applications of computer-generated faces in AI. Although current technology can eliminate differences in the physical appearance of computer-generated and natural faces, we argue that this may not be sufficient to eliminate biases against faces believed to be artificial, or artificial agents as a whole. In order to do so, social cognitive processes should be targeted that underlie how humans perceive trustworthiness in faces in light of prior attitudes and beliefs they hold about said faces. We emphasize the importance of distinguishing between technology-oriented and psychological-oriented inquiries in this emergent literature, as our findings strongly suggest that the perception of social attributes in faces is not solely driven by perceptual features of the

stimuli, but also, if not mainly, by higher level categorization processes capable of tainting perception.

References

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability*, *11*(1), 189. https://doi.org/10.3390/su11010189

Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior*, *52*, 331–337. https://doi.org/10.1016/j.chb.2015.06.018

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

Balas, B., & Tonsager, C. (2014). Face animacy is not all in the eyes: Evidence from contrast chimeras. *Perception*, *43*(5), 355–367. https://doi.org/10.1068/p7696

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/j.1467-9280.2007.01964.x

Billard, A., Robins, B., Nadel, J., & Dautenhahn, K. (2007). Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. *Assistive Technology: The Official Journal of RESNA*, *19*(1), 37–49. https://doi.org/10.1080/10400435.2007.10131864

Birkás, B., Dzhelyova, M., Lábadi, B., Bereczkei, T., & Perrett, D. I. (2014). Cross-cultural perception of trustworthiness: The effect of ethnicity features on evaluation of faces' observed trustworthiness across four samples. *Personality and Individual Differences*, *69*, 56–61. https://doi.org/10.1016/j.paid.2014.05.012

Burke, D., Nolan, C., Hayward, W. G., Russell, R., & Sulikowski, D. (2013). Is there an own-race preference in attractiveness? *Evolutionary Psychology: An International Journal of Evolutionary Approaches to Psychology and Behavior*, *11*(4), 855–872.

Crookes, K., Ewing, L., Gildenhuys, J., Kloth, N., Hayward, W. G., Oxner, M., Pond, S., & Rhodes, G. (2015). How well do computer-generated faces tap face expertise? *PLoS ONE*, *10*(11), e0141353. https://doi.org/10.1371/journal.pone.0141353

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). “Their ideas of beauty are, on the whole, the same as ours”: Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, *68*(2), 261–279. https://doi.org/10.1037/0022-3514.68.2.261

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, *96*(1), 108–116. http://blockqai.com/wp-content/uploads/2021/01/analytics-hbr-ai-for-the-real-world.pdf

De Houwer, J. (2006). What are implicit measures and why are we using them? In R. Wiers & A. Stacy, *Handbook of Implicit Cognition and Addiction* (pp. 11–28). SAGE Publications, Inc. https://doi.org/10.4135/9781412976237.n2

De Rosario-Martinez, H. (2015). *phia: Post-Hoc Interaction Analysis* (0.2-1) [R]. https://CRAN.R-project.org/package=phia

Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour*, *1*(1), 1–6. https://doi.org/10.1038/s41562-016-0001

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. https://doi.org/10.1037/a0022327

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1–20. https://doi.org/10.1037/0022-3514.90.1.1

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism: Clinical and Experimental*, *69S*, S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot? *Ergonomics in Design*, *19*(3), 24–29. https://doi.org/10.1177/1064804611415045

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529. https://doi.org/10.1037/pspa0000090

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Jackson, J. L. (2018). The non-performativity of implicit bias training. *Radical Teacher*, *112*, 46–54. https://doi.org/10.5195/rt.2018.497

Jones, D. (1995). Sexual selection, physical attractiveness, and facial neoteny: Cross-cultural evidence and implications. *Current Anthropology*, *36*(5), 723–748. https://doi.org/10.1086/204427

Kätsyri, J. (2018). Those virtual people all look the same to me: Computer-rendered faces elicit a higher false alarm rate than real human faces in a recognition memory task. *Frontiers in Psychology*, *9*(AUG). https://doi.org/10.3389/fpsyg.2018.01362

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marshburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., … Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology. General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

Langlois, J. H., Roggman, L. A., Casey, R. J., Ritter, J. M., Rieser-Danner, L. A., & Jenkins, V. Y. (1987). Infant preferences for attractive faces: Rudiments of a stereotype? *Developmental Psychology*, *23*(3), 363–369. https://doi.org/10.1037/0012-1649.23.3.363

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Levin, D. T. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1364–1382. https://doi.org/10.1037/0278-7393.22.6.1364

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1638–1659. https://doi.org/10.1098/rstb.2010.0404

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*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

MacLin, O. H., & Malpass, R. S. (2001). Racial categorization of faces: The ambiguous race face effect. *Psychology, Public Policy, and Law*, *7*(1), 98–118. https://doi.org/10.1037/1076-8971.7.1.98

MacLin, O. H., & Malpass, R. S. (2003). The ambiguous-race face illusion. *Perception*, *32*(2), 249–252. https://doi.org/10.1068/p5046

Matarić, M., Tapus, A., Winstein, C., & Eriksson, J. (2009). Socially assistive robotics for stroke and mild TBI rehabilitation. *Studies in Health Technology and Informatics*, *145*, 249–262.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Meissner, C. A., Brigham, J. C., & Butz, D. A. (2005). Memory for own- and other-race faces: A dual-process approach. *Applied Cognitive Psychology*, *19*(5), 545–567. https://doi.org/10.1002/acp.1097

Ng, W.-J., & Lindsay, R. C. L. (1994). Cross-race facial recognition: Failure of the contact hypothesis. *Journal of Cross-Cultural Psychology*, *25*(2), 217–232. https://doi.org/10.1177/0022022194252004

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(8). https://doi.org/10.1073/pnas.2120481119

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., Zoll, C., & Hall, L. (2004). Caring for agents and agents that care: Building empathic relations with synthetic agents. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, 194–201.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051–1068. https://doi.org/10.1037/0022-3514.42.6.1051

Quattrone, G. A., & Jones, E. E. (1980). The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, *38*(1), 141–152. https://doi.org/10.1037/0022-3514.38.1.141

Ramos, T., Oliveira, M., Santos, A. S., Garcia-Marques, L., & Carneiro, P. (2016). Evaluating young and old faces on social dimensions: Trustworthiness and dominance. *Psicologica*, *37*(2).

Rhodes, G., Simmons, L. W., & Peters, M. (2005). Attractiveness and sexual behavior: Does attractiveness enhance mating success? *Evolution and Human Behavior*, *26*(2), 186–201. https://doi.org/10.1016/j.evolhumbehav.2004.08.014

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001). Attractiveness of facial averageness and symmetry in non-western cultures: In search of biologically based standards of beauty. *Perception*, *30*(5), 611–625. https://doi.org/10.1068/p3123

Rodin, M. J. (1987). Who is memorable to whom: A study of cognitive disregard. *Social Cognition*, *5*(2), 144–165. https://doi.org/10.1521/soco.1987.5.2.144

Salvia, J., Sheare, J. B., & Algozzine, B. (1975). Facial attractiveness and personal-social development. *Journal of Abnormal Child Psychology*, *3*(3), 171–178. https://doi.org/10.1007/BF00916748

Singmann, H., Bolker, B., Westfall, J., Aust, F., & S. Ben-Shachar, M. (2021). *afex: Analysis of Factorial Experiments* (1.0-1) [R]. https://CRAN.R-project.org/package=afex

Sporer, S. L. (2001). The cross-race effect: Beyond recognition of faces in the laboratory. *Psychology, Public Policy, and Law*, *7*(1), 170–200. https://doi.org/10.1037/1076-8971.7.1.170

Stanley, D. A., Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., & Phelps, E. A. (2012). Race and reputation: Perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1589), 744–753. https://doi.org/10.1098/rstb.2011.0300

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*(1), 1–39. https://doi.org/10.1146/annurev.ps.33.020182.000245

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of Intergroup Relations* (2nd ed.). Hall Publishers.

Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (2022). *On the realness of people who do not exist: The social processing of artificial faces*. PsyArXiv. https://doi.org/10.31234/osf.io/dnk9x

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *43*(2), 161–204. https://doi.org/10.1080/14640749108400966

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, *69*(10), 1996–2019. https://doi.org/10.1080/17470218.2014.990392

Van Dessel, P., De Houwer, J., Gast, A., Roets, A., & Smith, C. T. (2020). On the effectiveness of approach-avoidance instructions and training for changing evaluations of social groups. *Journal of Personality and Social Psychology*, *119*(2), e1–e14. https://doi.org/10.1037/pspa0000189

Van Dessel, P., De Houwer, J., Gast, A., & Tucker Smith, C. (2015). Instruction-based approach-avoidance effects: Changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology*, *62*(3), 161–169. https://doi.org/10.1027/1618-3169/a000282

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wirtz, J., Patterson, P. G., Kunz, W. H., Gruber, T., Lu, V. N., Paluch, S., & Martins, A. (2018). Brave new world: Service robots in the frontline. *Journal of Service Management*, *29*(5), 907–931. https://doi.org/10.1108/JOSM-04-2018-0119

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, *35*(2), 151–175. https://doi.org/10.1037/0003-066X.35.2.151