UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Liefooghe, B., Oliveira, M., Leisten, L. M., Hoogers, E., Aarts, H., & Hortensius, R. (2023). Are Natural Faces Merely Labelled as Artificial Trusted Less? *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.73066

Social Psychology

Are Natural Faces Merely Labelled as Artificial Trusted Less?

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

[1] Department of Psychology, Utrecht University, Utrecht, Netherlands, [2] Radboud University, Nijmegen, Netherlands

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

https://doi.org/10.1525/collabra.73066

Collabra: Psychology

Vol. 9, Issue 1, 2023

Artificial intelligence increasingly plays a crucial role in daily life. 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. In five experiments (total $N = 867$), we tested whether natural faces that are merely labelled to be artificial are also trusted less. A meta-analysis of all five experiments suggested that natural faces merely labeled as being artificial were judged to be less trustworthy. This bias did not depend on the degree of trustworthiness and attractiveness of the faces (Experiments 1-3). It was not modulated by changing raters' attitude towards artificial intelligence (Experiments 2-3) or by information communicated by the faces (Experiment 4). We also did not observe differences in recall performance between faces labelled as artificial or natural (Experiment 3). When participants only judged one type of face (i.e., either labelled as artificial or natural), the difference in trustworthiness judgments was eliminated (Experiment 5) suggesting that the contrast between the natural and artificial categories in the same task promoted the labelling effect. We conclude that faces that are merely labelled to be artificial are trusted less in situations that also include faces labelled to be real. We propose that understanding and changing social evaluations towards artificial intelligence goes beyond eliminating physical differences between artificial and natural entities.

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). These 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 (e.g., 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 can pertain even for artificial faces that are undistinguishable from real faces.

People are generally able to rapidly form impressions about someone's trustworthiness based on their facial appearance alone (B. C. Jones et al., 2021; Oosterhof & Todorov, 2008; Todorov et al., 2009), regardless of whether or not these impressions are able to predict the behavior of the judged person (for a review see Todorov et al., 2015, pp. 531–535). These appearance-driven judgments are influenced by the degree to which we are exposed to particular distributions of facial features in our environment (e.g., Dotsch et al., 2016; W.-J. Ng & Lindsay, 1994). Pre-

a 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

vious 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). From the perspective of a human observer, artificial faces may thus 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 bears 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. Current technology enables to render faces that look completely realistic (see for instance, https://thispersondoesnotexist.com/). In addition, 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 remembered 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 the proposal 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 and this 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, however, always presented with real natural faces.

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 a 7-point Likert scale. In addition, facial judgments of attractiveness were assessed. Attractiveness judgments 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). In other words, the inclusion of attractiveness judgments helps assess the extent to which any effect of labelling is exclusive to a trustworthiness dimension, as opposed to any other dimension strongly related to general valence. 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: < 18 years: n = 1; 18-24 years: n = 27; 25-34 years: n = 10; 45-54 years: n = 4; 55-64 years: n = 17; and > 85 years: n = 1; no information about participant ethnicity). All these participants were included in the analyses. Participants belonged to the social network of the fourth author of the study[1] and participated for free. This sample size ($N = 60$) allows to detect an effect as small as $d = 0.465$, with 80% power and alpha = .05, for a within-subjects design (with 24 target stimuli nested within condition). The sample size is thus sufficiently large for our research purpose.

1 Experiments 1 and 2 were part of the Bachelor Thesis of the fourth author.

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; SE = 0.38) and high attractive faces (M = 4.15; SE = 0.23); high trustworthy (M = 4.17; SE = 0.24) and low attractive faces (M = 3.00; SE = 0.11); low trustworthy (M = 2.25; SE = 0.26) and high attractive (M = 3.68; SE = 0.44); low trustworthy (M = 1.93; SE = 0.19) and low attractive (M = 2.60; SE = 0.15). To ensure that participants rated the faces proper, they were cropped in an 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. 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). Face categories were randomized within a block such that any face category could change to another at every trial. 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. At the beginning of each block, 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). See also an example illustrating the task in Figure 1).

Data analysis. For the analysis of all experiments in the present study, linear mixed models were used as implemented in the packages 'lme4' (Bates et al., 2014) and 'lmerTest' (Kuznetsova et al., 2017). Estimated marginal means were calculated with the package 'emmeans' (Lenth et al., 2018). Facial Attractiveness (High vs. Low), Facial Trustworthiness (High vs. Low) and Label (Natural, Computer-generated) were fixed effects and effect coded. The significance of fixed effects and their interactions was assessed with F-tests using Satterwhaite's method for estimating the degrees of freedom of the denominator (Luke, 2016). Random intercepts were included for the grouping variable Participant and the grouping variable Face. In addition, random slopes were also added and were allowed to correlate. We thus started with the maximal random-effects structure (Barr et al., 2013). For the trustworthiness ratings, this resulted in a singular model and the random-effects structure was simplified following Bates et al. (2014). Principal components analysis of the estimated covariance matrices for the random effects indicated that only six out of eight dimensions were sufficient to account for the full variance. In a second step, the zero-correlation parameter model was fitted. This model remained singular. The random-effects structure was further simplified by removing the smallest variance components and performing Likelihood Ratio Tests. These steps resulted in the following random-effect structure: [Label + Facial Attractiveness + (Label x Facial Trustworthiness) + (Label x Facial Attractiveness x Facial Trustworthiness) || Participant] + (1|Face). For the attractiveness ratings the full model was also singular, and the zero-correlation parameter model was used: (Label x Facial Attractiveness x Facial Trustworthiness || Participant) + (Label || Face).

Results

Participants' overall belief that the computed-generated faces were actually generated by a computer was 5.78 (SE = 1.18). This was significantly higher than the middle of the Likert scale, $t(60) = 11.93$, $p < .001$. The percentage of participants distributed across the belief scale is shown in Figure 2.

Trustworthiness. Faces labelled to be computer-generated (M = 3.82; SE = 0.13) were rated to be less trustworthy compared to faces labelled to be natural (M = 3.97; SE = 0.13), $F(1,59.10) = 6.35$, $p < .05$. High trustworthy faces (M = 4.36; SE = 0.15) were rated to be more trustworthy than low trustworthy faces (M = 3.43; SE = 0.15), $F(1, 21.48) = 25.94$, $p < .001$. High attractive faces (M = 4.18; SE = 0.15) were rated to be more trustworthy compared to low attractive faces (M = 3.61; SE = 0.15), $F(1, 21.98) = 9.63$, $p = .005$ (see Figure 3). The interactions were not significant, all $Fs < 1$.

Attractiveness. Mean attractiveness ratings did not differ reliably between both labels (natural: M = 3.00; SE = 0.14; computer-generated: M = 3.07; SE = 0.14), $F(1, 24.34) = 1.54$, $p = .227$. High attractive faces (M = 3.94; SE = 0.18) were rated to be more attractive compared to low attractive faces (M = 2.13; SE = 0.18), $F(1, 26.15) = 64.00$, $p < .001$. High trustworthy faces (M = 3.44; SE = 0.17) were also rated to be more attractive than low trustworthy faces (M = 2.63; SE = 0.17), $F(1, 20.95) = 14.34$, $p = .001$. The interactions were not significant. The largest F-value of an interaction was observed for the interaction between attractiveness and trustworthiness, $F(1, 20.06) = 1.54$, $p = .228$.

Figure 1. Illustration of the Rating Task and Stimuli of Experiments 1, 2, 3, and 5

Note. Panel A: Example of the face rating task in Experiments 1, 2, 3, and 5. Only one face was shown at each trial, paired with only one of the labels. Panel B: Example of stimuli extracted from Experiment 3 that were rated, on average, as high (top row) or low (bottom row) in trustworthiness.

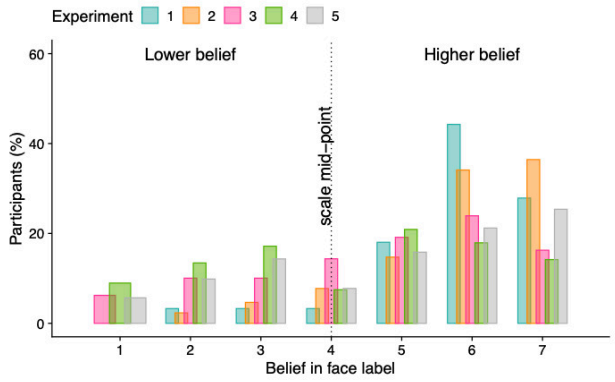

Figure 2. Distributions of Believability Ratings for All Experiments

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.

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

Participants. A sample of 130 participants was recruited through social media (100 female, 29 male; age categories: < 18 years: n = 8; 18-24 years: n = 86; 25-34 years: n = 23; 35-44 years: n = 1; 45-54 years: n = 8; 55-64 years: n = 2; and 65-74 years: n = 2; no information about participant ethnicity). A gift voucher of 25 euros 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.

Materials. Materials and procedure were identical to Experiment 1, for the exception that negative or positive cover story was added before the rating task. Following the rating of the faces, the extent to which participants believed that the computer-generated faces were 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.

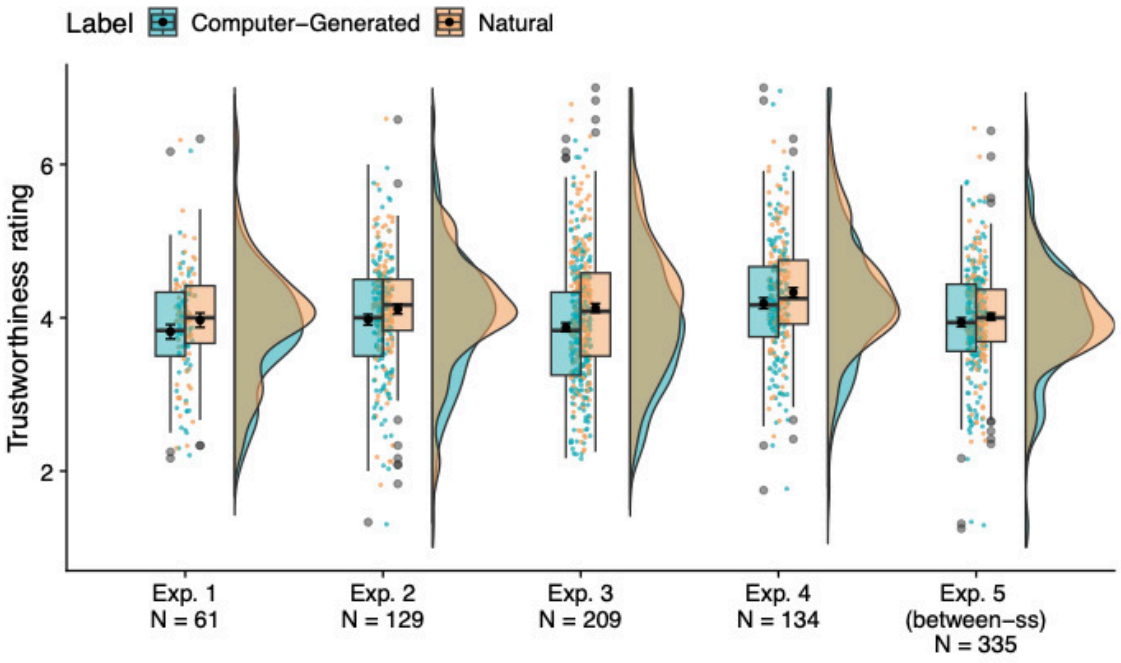

Figure 3. Raincloud Plots of the Main Effects of Label in All Experiments

Data analysis. Story (Positive, Negative), Facial Attractiveness (High, Low), Facial Trustworthiness (High, Low), Label (Natural, Computer-generated) were effect coded as fixed effects. The maximal random-effects structure could be estimated for both dependent variables: (Label x Facial Attractiveness x Facial Trustworthiness | Participant) + (Label x Story | Face).

Results

Participants' overall belief that the computer-generated faces were actually generated by a computer was 5.83 (*SE* = 1.26), which was significantly higher than the middle of the Likert scale, $t(128) = 16.45$, $p < .001$ (see Figure 2). In addition, participants were significantly more positive towards artificial intelligence in the condition in which a positive cover story was used ($M = 4.27$; $SE = 0.16$) compared to the condition in which a negative cover story was used ($M = 3.80$; $SE = 0.15$), $t(127) = 2.13$, $p = .035$, $d = -0.38$, 95% CI [-0.73, - 0.02].

Trustworthiness. Similar to Experiment 1, faces labelled to be computed-generated ($M = 3.98$; $SE = 0.11$) were rated to be less trustworthy compared to faces labelled to be natural ($M = 4.11$; $SE = 0.10$). However, the main effect of Label failed to be significant, $F(1, 91.56) = 3.35$, $p = .070$ (see Figure 3). There was no main effect of Cover Story, $F < 1$ (Positive: $M = 4.03$, $SE = 0.11$; Negative: $M = 4.06$, $SE = 0.12$). Cover Story did not interact with Label, $F < 1$.

High trustworthy faces ($M = 4.51$; $SE = 0.13$) were rated to be more trustworthy than low trustworthy faces ($M = 3.58$; $SE = 0.13$), $F(1, 22.29) = 31.68$, $p < .001$. High attractive faces ($M = 4.23$; $SE = 0.13$) were rated to be more trustworthy compared to low attractive faces ($M = 3.86$; $SE = 0.13$), $F(1, 21.25) = 5.17$, $p = .033$. None of the interactions was significant. The largest *F*-value of an interaction was observed for the interaction between Label and Facial Trustworthiness, $F(1, 43.33) = 1.76$, $p = .192$.

Attractiveness. Attractiveness ratings did not differ reliably between faces labelled as natural ($M = 3.31$; $SE = 0.12$) and faces labelled as computer-generated ($M = 3.24$; $SE = 0.12$), $F(1, 57.09) = 1.98$, $p = .165$. Ratings did not differ significantly as a function of Story, $F < 1$ (Negative: $M = 3.29$, $SE = 0.14$; Positive: $M = 3.27$, $SE = 0.13$). Story did not interact significantly with Label, $F < 1$.

High attractive faces ($M = 4.11$; $SE = 0.15$) were rated to be more attractive compared to low attractive faces ($M = 2.45$; $SE = 0.16$), $F(1, 24.64) = 65.99$, $p < .001$. High trustworthy faces ($M = 3.61$; $SE = 0.15$) were also rated to be more attractive than low trustworthy faces ($M = 2.94$; $SE = 0.15$), $F(1, 20.41) = 11.58$, $p = .003$. The remaining interactions were all not signification, all $Fs < 1$.

Discussion

Faces labelled as computer-generated were rated as being less trustworthy compared to faces labelled as natural. However, this effect was numerically present but not significant. 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' attitude towards AI, but not the difference in perceived trustworthiness between faces labelled as computer-generated and faces labelled as natural. For the attractiveness ratings no reliable difference was observed between both types of faces.

Taken together, the results of Experiment 2 are in line with the results of Experiment 1 and again suggest that faces believed to be computer-generated may be considered as 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 Experiments 1 & 2 with a more optimal design. First, more variation in the faces used was allowed by sampling faces randomly for each participant from pre-selected sets of faces. A more stringent protocol was used to create these sets (see below). Second, power calculations were now geared to specifically test an interaction between the effect of label and the background story. Third, we tested 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. An initial sample of 210 participants was recruited via Prolific Academic (www.prolific.co). The sample size was determined by using power simulations, which are presented in the pre-registration protocol (https://osf.io/w4bca). 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). Data about participants' self-reported ethnicity was retrieved from Prolific Academic and revealed that most participants identified themselves as White (White = 70.4%, Black = 22.8%, Mixed = 4.4%, Asian = 1%, Other = 1.5%; according to the ethnicity categories provided by Prolific). 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. Both scales ranged from 1 (Not at all) to 7 (Extremely). 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.

Data analysis. 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 used linear mixed models to analyze d' and c as a function of Story and Label.

For the ratings, linear mixed models were constructed and evaluated as in the previous experiments. For trustworthiness, the maximal random-effects model was singular and simplified, resulting in the following random-effects structure: (Label + Facial Attractiveness + Facial Trustworthiness || Participant) + [Story + (Story x Label) || Face]. For attractiveness the random-effect structure was as follows: (Label + Facial Attractiveness + Facial Trustworthiness || Participant) + (1| Face).

These analyses, following a suggestion by a reviewer, deviated from the originally pre-registered ANOVA analyses and improve upon them by taking into account more sources of variability in the data. The results of the pre-registered ANOVAs converged entirely with the results of the linear mixed models.

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, $t(208) = 5.38$, $p < .001$ (see Figure 2). Similar to Experiment 2, participants were significantly more positive towards artificial intelligence in the condition in which a positive cover story was used (M = 4.82, SE = 0.12) compared to the condition in which a negative cover story was used (M = 3.77; SE = 0.14), $t(207) = -5.67$, $p < .001$, d = 0.79, 95% CI [0.74, 0.83].

Trustworthiness. The interaction between Story and Label was not significant, $F < 1$. Neither was the main effect of Story, $F < 1$. The main effect of Label was significant, $F(1, 208.00) = 35.17$, $p < .001$ (see Figure 3). Faces labeled as computer-generated were rated as less trustworthy (M = 3.88, SE = 0.07) than faces labeled as natural (M = 4.12, SE = 0.07). Finally, trustworthy faces were rated as more trustworthy (M = 4.40, SE = 0.08) than untrustworthy faces (M = 3.61, SE = 0.08), $F(1, 111.4) = 84.62$, $p < .001$, and attractive faces were rated as more trustworthy (M = 4.22, SE = 0.08) than unattractive faces (M = 3.79, SE = 0.08), $F(1, 94.00) = 26.94$, $p < .001$.

The Interaction between Face Trustworthiness and Face Attractiveness was also significant, $F(1, 88.50) = 6.59$, p = .012. The difference between trustworthy and untrustworthy faces was more pronounced for unattractive faces (M_{diff} = 0.99) than for attractive faces (M_{diff} = 0.59). The three-way interaction between Story, Face Trustworthiness, and Face Attractiveness was also significant, $F(1, 78.90) = 4.20$, p = .044. This interaction was not readily interpretable.

Attractiveness. There was no interaction between Story and Label nor a main effect of Story, both Fs < 1. The main effect of Label was significant, $F(1, 205.20) = 16.16$, $p < .001$, 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).

Finally, attractive faces were rated as more attractive (M = 4.28, SE = 0.06) than unattractive faces (M = 3.04, SE = 0.06), $F(101.50) = 116.87$, $p < .001$, and trustworthy faces were rated as more attractive (M = 3.89, SE = 0.06) than untrustworthy faces (M = 3.43, SE = 0.06), $F(1, 100.00) = 16.82$, $p < .001$. None of the interactions was significant. The largest F-value of an interaction was obtained for the four-way interaction, $F(1, 4104.90) = 1.28$, p = .259.

Memory task. Regarding sensitivity (d'), participants showed good ability to detect faces that had been previously shown (one-sided t-tests against zero: mean $d'_{\text{computer-generated}}$ = 0.97, SE = 0.04, $t(208) = 25.49$, $p < .001$; mean d'_{natural} = 0.94, SE = 0.04, $p < .001$, $t(208) = 25.86$, $p < .001$). Separate linear mixed models were conducted for sensitivity (d') and response bias). Both models included Story, Label, and their interaction as fixed effects. The random effects were specified to allow the intercepts and slopes of Label to vary by participant, and intercepts to vary by face stimulus. The mixed model for sensitivity[2] revealed no interaction between Story and Label $F(1, 275.34) = 1.03$, p = .359. There was also no main effect of Story, $F(1, 206.5) = 1.32$, p = .252. The effect of Label was significant, $F(1, 275.44) = 7.90$, $p < .001$. However, follow-up analyses of simple effects clarify revealed no difference between the mean sensitivity between faces previously labeled as natural and that of faces previously labeled as computer-generated, since these means were the same ($M_{d'}$ = 0.45 for both label conditions, p = .992). The significant effect was instead mainly driven by the difference between distractor faces (i.e., faces not previously presented in the main rat-

2 Following a suggestion by a reviewer, we conducted an alternative, non-preregistered, generalized mixed model analysis to account for the randomizing of face stimulus identities between subjects, and for the stimulus sampling. The model was specified similarly to the d' linear mixed models, with the exception that accuracy in the memory task was now the dependent variable (coded as 1= correctly identified stimulus as old or new, 0 = incorrectly identified stimulus as old or new). The results, reported on the logit scale, entirely converged with those obtained for the d' analysis, showing no significant difference in accuracy between stimuli that had been labeled as natural and computer-generated ($b_{Computer\text{-}Generated\_Label}$ = -0.08, SE = 0.16, p = .63, Label reference level: natural; odds ratio (natural / computer-generated) = 0.95, SE = 0.11, p = .91).

ing task) and faces previously labelled as natural (M_{diff} = 0.08, p < .001) or as computer-generated (M_{diff} = 0.07, p < .001). The response bias analysis revealed an equal tendency to commit false alarms for faces previously labeled as computer-generated (c = -0.23, SE = 0.01) or as natural (c = -0.23, SE = 0.01). Although the Label effect emerged as the only significant one, $F(1, 285.43.5) = 1268.25$, p < .001, its significance was mainly driven by the difference between distractor stimuli and stimuli that had been previously labeled (both differences M_{diff} = -0.41, p < .001), and not by any difference between stimuli previously labelled as natural and stimuli labelled as computer-generated (M_{diff} = -0.002, p = .847).

Discussion

In line with 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 judgment biases despite of a 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 findings of Experiments 2-3 thus suggest that merely labelling a face as being computer-generated is sufficient to mimic the effect of a well-established social outgroup (Jackson, 2018; Lai et al., 2016; Van Dessel et al., 2020 for a discussion), which is difficult to modulate by simply adding contextual information. However, an important behavioral marker for outgroup effects is that outgroup faces are remembered less well and the results of the memory task in Experiment 3 indicated that faces were remembered equally well regardless of the label they were paired with at initial exposure. In contrast, Bernstein and colleagues (2007) also used arbitrary social categorization and did observe impaired memory for faces arbitrarily categorized as an outgroup, while controlling for differences in perceptual face expertise. Furthermore, previous studies observed a memory advantage for natural faces compared to visibly computer-generated faces (Balas & Pacella, 2015; Crookes et al., 2015; Kätsyri, 2018).

An additional element suggesting that the label effect is not based on the creation of an outgroup relates to the observation 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.

Interestingly, apart from Experiment 3, we did not observe reliable differences between faces labelled as computer-generated or as natural for attractiveness ratings in Experiments 1-3. Although attractiveness may be processed differently between real and synthetic faces (Balas et al., 2018), evidence supporting clear differences in attractiveness between in- and outgroup faces is generally mixed (Burke et al., 2013; Cunningham et al., 1995; D. 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. Because the attractiveness ratings were not our prime concern, we did not administer these ratings in the follow-up experiments.

Taken together, Experiments 1-3 do not offer strong support for our initial hypothesis that faces labelled to be artificial are perceived to be less trustworthy because they are represented as an outgroup. Accordingly, we conducted two additional experiments that further investigate the label effect by testing its boundary conditions.

Experiment 4

Research focusing on cooperative trust between humans and algorithms has indicated that when humans play a trust game against an algorithm they are less willing to cooperate (Crandall et al., 2018; Ishowo-Oloko et al., 2019; Kiesler et al., 1996; Miwa & Terai, 2012; Y.-L. Ng, 2022; Oksanen et al., 2020). Interestingly, the degree to which the algorithm is programmed to be cooperative does not moderate this bias (Ishowo-Oloko et al., 2019; Miwa & Terai, 2012). Fueled by this finding, Experiment 4 further investigated the label effect when using short video clips of faces verbally expressing information to the perceiver, therefore including both facial appearance and voice across a brief period of time—akin to a "thin slice of behavior" (Ambady & Rosenthal, 1992). Importantly, we manipulated the accuracy of the information communicated by the faces and examined its impact on the ratings of facial trustworthiness. Each face said aloud the following message: "It's ten minutes to two o'clock." The faces were presented together with a digital clock, which could indicate the same or a different time. We assumed that the communication of inaccurate information might affect perceptions of trustworthiness (e.g., by sounding deceitful, or as unreliable). As such, we could test if the label effect is modulated by the 'behavior' of the faces and further understand the inferences participants make when rating faces. Participants may only rate the faces proper. Accordingly, the correctness of the messages should not interact with the label effect. Alternatively, participants may try to make some inferences or hypotheses about the agent and use these to rate the trustworthiness. If the label effect is mediated by such inferences about the agent behind the face, then the label effect may interact with the correctness of the message. We had no specific hypothesis about the directionality of this interaction.

A final change of Experiment 4 compared to the previous experiments is that we used three different wordings per label category. The natural faces could be labelled by the words "natural", "human", and "real". The artificial faces were labelled with the words "artificial", "computer-generated", and "synthetic". Per participant random pairs of these words were selected from each category. As such, we could rule out the possibility that the labelling effect found in the previous experiments is being driven by the specific words ('computer-generated', 'natural') that were used.

Method

Participants. This experiment was preregistered. An initial sample of 135 participants was recruited via Prolific Academic (www.prolific.co). The sample size was determined based on power simulations that took the effects sizes of the previous experiments into account. For detailed discussion see the preregistration protocol (https://osf.io/h6tk2). Due to the language spoken in the stimuli, we only recruited participants who were fluent in French according to the Prolific Academic database. We did not filter for participation in any of our previous experiments. However, we were able to verify if any unique participant IDs overlapped across Experiments 3, 4, and 5, and found no duplicates. This suggests that all participants had not been previously exposed to the experimental setup. One participant was excluded from the analysis due to zero variability in the ratings. The final sample included in the analyses consisted of 134 participants (59 female, 68 male; 4 non-binary/third gender; other gender = 2, undisclosed gender = 1; median age (IQR): 29 (14), age range: 19-67). The final sample was thus one participant short of the pre-registered sample size of 135.). Data about participants' self-reported ethnicity was retrieved from Prolific Academic and revealed that most participants identified themselves as White (White = 78.5%, Black = 8.1%, Mixed = 8.1%, Asian = 3%, Other = 2.2%; according to the ethnicity categories provided by Prolific). All participants, including the excluded and additional ones, were compensated with £1.60 according to an hourly rate of £8.00.

Materials. The stimuli were extracted from the Geneva Faces and Voices (GEFAV) database (Ferdenzi et al., 2015), and consisted of videos of female and male faces merged with their respective audio recordings (i.e., voices) of different individuals verbally expressing the same message out loud (viz. "Bonjour. Il est deux heures moins dix."). Please note that information about face ethnicity is not offered by the database. All faces and voices are described as belonging to French-speaking natives of European origin. The videos and audio files had to be merged as they are originally decoupled in the GEFAV database. The individuals in the videos displayed a neutral facial expression and their faces were cropped such that only the internal face features remained visible (i.e., no hair). The GEFAV database includes ratings of perceived facial trustworthiness for the videos and perceived voice trustworthiness for the audio files. The voice ratings include ratings by human judges of either sex (male or female), or an average of both. We chose to only use the (averaged) voice ratings generated both sexes. For each identity in a video, we computed an index of average face and voice trustworthiness and used this index to classify the stimuli into Low, Medium, and High average face and voice trustworthiness (i.e., terciles of the distribution of averaged ratings). We then created a stimulus pool including the stimuli classified as Medium. The videos were then pretested in regard to the quality of the synchronization between the video and the audio recording (as these were manually merged) on a 4-point scale ranging from 1 (Not at all synchronized) to 4 (Perfectly synchronized). Videos with mean ratings below 3 on the pretest were dropped from the stimulus pool. Based on the pretest results, the final stimulus pool included a total of 30 videos. For each participant, 24 stimuli were randomly drawn from this pool. The final stimulus set of 30 faces was imbalanced in terms of face sex with a higher proportion of female faces (11 male, 19 female), thus preventing us from performing the ideal counterbalancing without losing stimuli. This means that, on average, participants would be exposed to more female identities compared to male identities. This imbalance was, however, approximately balanced between the label categories (Natural category: 1006 female vs. 602 male trials; Non-natural category: 1045 female vs. 563 male trials).

Two categories of labels were used to indicate the nature of the stimuli: natural and non-natural. There were three words for the label per category. The words used to represent the natural category were: "natural", "human", and "real". The words used to represent the non-natural category were: "artificial", "computer-generated", and "synthetic". The resulting unique pairings of a natural word label with a non-natural word label were counter-balanced across participants. Thus, for example, while some participants were exposed to "human" versus "artificial", others were exposed to "real" versus "computer-generated".

The accuracy of the message verbally expressed by the face in each video was manipulated by pairing the video with information that either matched or mismatched with the message expressed by the face in the video. Specifically, each video was paired with an image of an alarm clock displaying a time that wither matched or mismatched with the time communicated by the face in the video. The time communicated by the faces was always the same (translation from French: "Good morning. It's ten minutes to two o'clock.") and the time displayed in the clock images varied across the inaccurate trials.

The main dependent variable of this study was the measure the perceived trustworthiness of each face stimulus on a scale ranging from 1 to 7 (Not at all to Extremely).

On a separate screen after each video paired with a clock time, we asked the participants to indicate if the time expressed by the person in the video matched with the time displayed in the clock (yes or no response, coded as 1 and 0 respectively). Again, we asked participants to what extent they believed that the stimuli had an artificial origin using a 7-point scale ranging from 1 (Not at all) to 7 (Extremely).

Two additional questions at the end of the experimental session measured a participant's self-reported attitude towards stimuli that are artificially generated/non-natural,

and their general attitude towards artificial intelligence. Both questions used a 7-point scale ranging from 1 (Extremely negative) to 7 (Extremely positive).

Procedure. An illustration of the experiment setup is shown in Figure 4. The experiment was conducted online using Qualtrics software. After consenting to participate in the study, participants were informed that the goal of the study was to investigate how people perceive the trustworthiness of people who were either natural or generated artificially. In the main task, participant first evaluated the trustworthiness of the person in the video, using the scale of 7-points ranging from 1 (Not at all) to 7 (Extremely) and subsequently indicated in the screen if the time expressed in the video matched the time displayed in the alarm clock, by clicking "Yes" or "No". A practice block with one trial—always showing the label representing the natural category word assigned to the participant—was completed before the main task. Before the practice block and before the main task, participants were asked to ensure that their audio was on so they could listen to the message in the video. For each participant, 25 videos were randomly drawn from the stimulus pool of 30 videos: one video was assigned to the practice block, and the remaining 24 videos were assigned to the main task. In each trial, the video was initiated automatically. After completing the main task, participants were asked to report the extent to which they believed in the nature of the videos (scale ranging from 1 – Not at all to 7 – Extremely), their general attitude towards AI and their attitude towards artificially generated stimuli (both measured on a scale ranging from 1 – Extremely negative to 7 – Extremely positive). Finally, they provided demographical information, were thanked, debriefed, and compensated.

Data analysis. Label Category (natural vs. non-natural) and Message Correctness (Correct vs. Incorrect) were effect coded fixed effects. Random intercepts were included for Participant and Face. Random slopes were also included. The maximal random-effect structure was singular and simplified: (Label x Message Correctness || Participant) + (1 | Face). These analyses, following a suggestion by a reviewer, deviated from the originally pre-registered ANOVA analyses and improve upon them by taking into account more sources of variability in the data. The results of the pre-registered ANOVAs converged entirely with the results of the linear mixed models.

Results

The main effect of Label Category was significant, $F(1, 133.32) = 8.91$, $p = .003$, indicating that faces labeled as computer-generated were rated as less trustworthy ($M = 4.18$, $SE = 0.09$) than their natural counterparts ($M = 4.33$, $SE = 0.09$) (see Figure 3). Faces indicating the correct time were also rated to be more trustworthy ($M = 5.08$, $SE = 0.12$) than faces indicating the incorrect time ($M = 3.43$, $SE = 0.12$), $F(1, 133.03) = 98.07$, $p < .001$. The two-way interaction was not significant, $F < 1$.

Participants' overall belief that the videos labelled as non-natural were actually generated artificially ($M = 4.28$, $SE = 0.17$) was again significantly higher than the middle of the scale, $t(133) = 1.71$, $p = .044$ (see Figure 2). Additional exploratory analyses[3] clarified that participants' overall attitude towards computer-generated face stimuli tended to be positive ($M = 4.27$, $SE = 0.11$; significantly different from middle of scale: $t(133) = 2.34$, $p = .010$) but did not moderate the two-way interaction, $F < 1$, or the label effect, $F(6, 127) = 1.98$, $p = .073$. The overall attitude towards AI also tended to be positive ($M = 4.85$, $SE = 0.12$; significantly different from middle of scale: $t(133) = 6.87$, $p < .001$) but also did not moderate the two-way interaction, $F < 1$, or the label effect, $F(6, 127) = 1.84$, $p = .096$.

Discussion

Experiment 4 replicates the results of the previous experiments. Faces merely labelled as being artificial are trusted less. Faces indicating the incorrect time were also considered to be less trustworthy compared to faces who indicated the correct time. However, both effects did not interact. The label effect does not seem to be modulated by the 'behavior' of the face. Such finding suggests that the label is effect is based on the judgment of the face rather than the judgment of the whole agent . Finally, in Experiment 4 different words were assigned to the two label categories. Nevertheless, we still replicated the label effect. This indicates that the label effect is not limited to the specific pairs of words used in Experiments 1-3.

Experiment 5

Across four experiments we have observed that the label effect is not modulated by the degree of trustworthiness and attractiveness of the faces, the raters' attitude towards artificial intelligence or by letting the faces making (in)correct assertions. In addition, no difference in memory performance was observed, a finding which is typically associated when faces are considered to be an outgroup. In Experiment 5 we tested whether the label effect could be driven by the demand to judge faces labelled as natural or labelled as computer-generated within the same task. The presence of both types of labels may trigger an implicit comparison between both categories of labels, which results in a contrast effect (see Ishowo-Oloko et al., 2019). Furthermore, it has been reported that trustworthiness

3 The models specified for these exploratory analyses slightly deviated from the pre-registered ones which specified attitude only as a covariate (additive term). Specifically, the reported analyses tested the three-way interaction of Label Category X Message Correctness X Attitude [towards CG or towards AI]. Any significant interaction between Attitude and any other effects would suggest that attitudes moderated the effect.

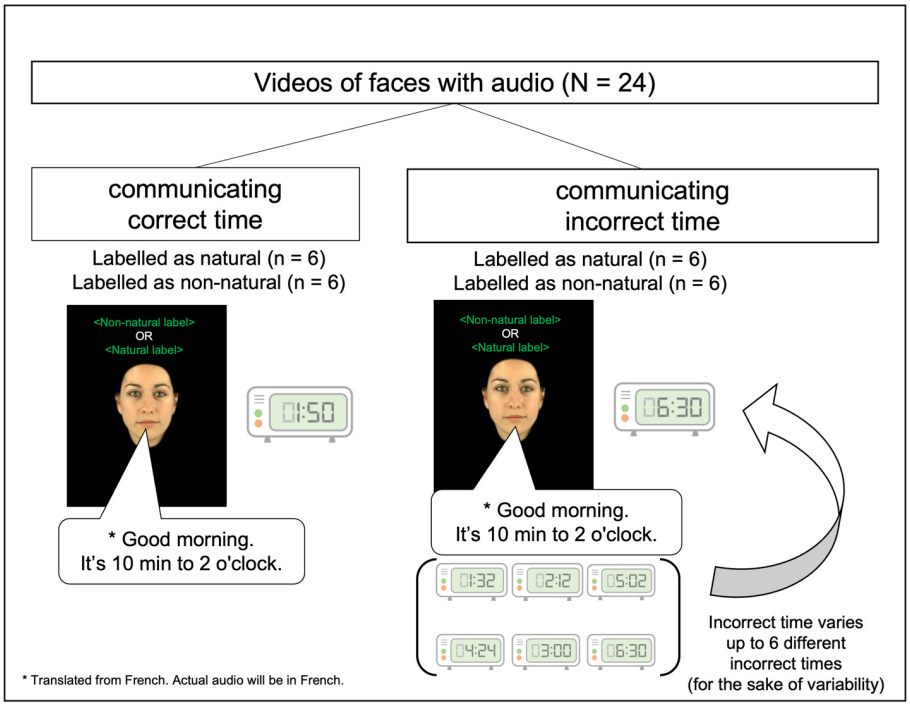

Figure 4. Experiment 4 Task Overview

judgments can be influenced through contrast effects (e.g., Schwarz & Bles, 1992).

We used static faces as in Experiments 1-3 and participants only rated one type of faces, thus either labelled as natural or as computer-generated. In addition, stimulus' ethnicity and sex were counterbalanced, and the set included only stimuli originally rated in the middle range of trustworthiness as in Experiment 4. For consistency, the instructions of the experiment only mentioned one possible category. The question was whether we could replicate the label effect under such conditions.

Method

Participants. This experiment was preregistered. An initial sample of 341 participants was recruited via Prolific Academic (www.prolific.co). The initial sample size was estimated based on power simulations presented in the preregistration protocol (https://osf.io/9kdu2). Six participants were dropped for not meeting our preregistered inclusion criterion for the variability in ratings (required *SD* > 0.5). A final round of recruitment aiming to replace the dropped participants resulted in final sample of 335 participants (161 female, 167 male; 6 non-binary/third gender; undisclosed gender = 1; median age (IQR): 25 (8), age range: 18-69), which overshoot our target sample size of 332 (see Power considerations). Prolific Academic participants who had participated in previous experiments of this paper were not allowed to participate in the study. Participants were fluent in English and reported being from 24 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 134 participants (59 female, 68 male; 4 non-binary/third gender; other gender = 2, undisclosed gender = 1; median age (IQR): 29 (14), age range: 19-67).). Data about participants' self-reported ethnicity was retrieved from Prolific Academic and revealed that most participants identified themselves as White (White = 63.6%, Black = 21.6%, Mixed = 9.8%, Asian = 2.1%, Other = 3%; according to the ethnicity categories provided by Prolific). All participants, including the excluded and additional ones, were compensated with £1.00 according to an hourly rate of £7.50.

Materials. A total of 48 different face identities were extracted from the CFD (Ma et al., 2015) before the experiment is conducted. Using the original face ratings (Ma et al., 2015), we first extracted the subset of stimuli who fell into the medium range (2nd tercile) of the distribution of trustworthiness ratings. Next, we counterbalanced the ethnicity of the faces in our set such that there were 12 faces per ethnicity. The version of the CFD (v1.0) that we used includes stimuli of individuals who self-identified as being of one out of four ethnicities in total (viz. Asian, Black, Latinx, or White). Within each ethnicity set, we counterbalanced the stimulus sex, resulting in six female and six male faces per ethnicity. The random sampling (seed = 4242) of face identities from the database was conducted separately for every subset of stimuli that resulted from crossing face ethnicity with face sex. For example, we first extracted the subset of stimuli that were both female and Asian, and subsequently randomly sampled six identities from this subset. This was repeated for all the remaining subsets. The resulting 48 face identities were then cropped in oval shape using webmorphR v0.1.1 (DeBruine, 2022) to maximize the comparability with the stimuli sets used in the previous Experiments. The code used to generate the stimulus sample and apply the cropping is available in the online repository.

The two words used to represent the natural and non-natural categories of the instructed origin of the stimuli were "Natural" and "Computer-generated", respectively.

The measures were in every way identical to the ones used in Experiments 1-3, with some exceptions. Only trustworthiness ratings were collected, and the attitudinal measures were identical to the ones used in Experiment 4 (attitude towards AI, or towards computer-generated faces). Due to the nature of the between-participants design, the believability question was tailored to the label condition, such that it asked whether the participants believed that the faces "were of real people" (natural condition) or "were actually generated by a computer" (non-natural condition) on a scale ranging from 1 (Not at all) to 7 (Extremely).

Procedure. The study was conducted online using Qualtrics software. The experimental procedure was similar to the one followed in Experiments 1-3, with some important differences, namely: participants only rated the trustworthiness of faces; the final block of questions included the tailored version of the believability question, and two attitudinal questions instead of one; there were 48 faces instead of 24; and there was no memory test. Participants were randomly assigned to either the natural or artificial label condition. The study instructions were tailored to each label condition. In the natural condition, participants were informed that the purpose of the study was to investigate how people perceive the trustworthiness of faces and were asked to rate the trustworthiness of 48 faces. In the artificial condition, participants were informed that the purpose of the study was to investigate how people perceive artificial faces and were shown a brief explanation about what computer-generated faces were and how current technology can render them as highly realistic, before receiving the same instructions to rate the stimuli. The same 48 faces were used in both label conditions. After completing the task, participants provided demographical information, were thanked, debriefed, and compensated.

Results

The main effect of Label was not significant, $F(1, 333.84) = 0.93$, $p = .334$. Trustworthiness ratings were similar for faces labelled as natural ($M = 4.01$, $SE = 0.09$) and faces labelled as non-natural ($M = 3.94$, $SE = 0.09$) (see also Figure 3).

In the artificial condition the mean believe that the faces were actually computer-generated was not significantly higher than the midpoint of the Likert scale ($M = 3.90$, $SE = 0.15$), $t(164) = 0.66$, $p = .745$. In contrast, mean believability was significantly higher than the midpoint of the Likert scale when participants were asked whether they believed the faces were from real people in the natural condition ($M = 5.73$, $SE = 0.11$), $t(169) = 15.57$, $p < .001$. Figure 2 shows how the participants are distributed across believability ratings regardless of Label condition. Positive attitudes towards computer-generated faces ($M = 4.22$, $SE = 0.01$) in the natural condition did not differ significantly from the attitude in the non-natural condition ($M = 4.00$, $SE = 0.02$), $t(333) = 1.54$, $p = .125$. Similarly, the general attitude towards artificial intelligence did not differ significantly between both conditions (Natural condition: $M = 5.06$, $SE = 0.01$; Non-Natural condition: $M = 4.84$, $SE = 0.01$), $t(333) = 1.61$, $p = .108$.

Discussion

In contrast to the previous experiments, we did not observe a label effect when faces were presented to two groups of participants, which only rated one type of face. This finding thus suggests that the label effect is possibly only present in contexts which require the judgment, or the presence of both faces labelled as natural and artificial. However, we note that in the artificial condition participants did not seem to strongly believe that the faces were actually generated by a computed. This lack in believability may also have driven our results and suggests that the manipulation was less successful as compared to the previous experiments. Another aspect that might have impacted the results is the improved counterbalancing of facial ethnicity and sex in the stimulus set. If the labelling effect is being driven by imbalances in ratings associated with stimulus features such as ethnicity and sex (e.g., Cook & Over, 2021; Sutherland et al., 2015), it is a possibility that counterbalancing those features would mitigate the effect. We further consider our results in the General Discussion.

Meta-analysis

We calculated the average effect size (Cohen's *d*) for the main effect of labelling across all five experiments. The Cohen's *d* values and respective 95% confidence intervals were derived from the *F* values and respective degrees of freedom reported above for these main effects, using the R package effectsize 0.6.0.1 (Ben-Shachar et al., 2020). The average effect size and forest plot were computed using the R package metaviz 0.3.1 (Kossmeier et al., 2020). The results are shown in Figure 5.

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 five experiments. The result of the meta-analysis of all five experiments suggests that natural faces merely labeled as being artificial were judged to be less trustworthy. This bias did not depend on the degree of trustworthiness and attractiveness of the faces (Experiments 1-3). This label effect was not modulated by changing raters' attitude towards artificial intelligence (Experiments 2-3) or by the correctness of messages communicated by the faces (Experiment 4). We also did not observe differences in recall performance be-

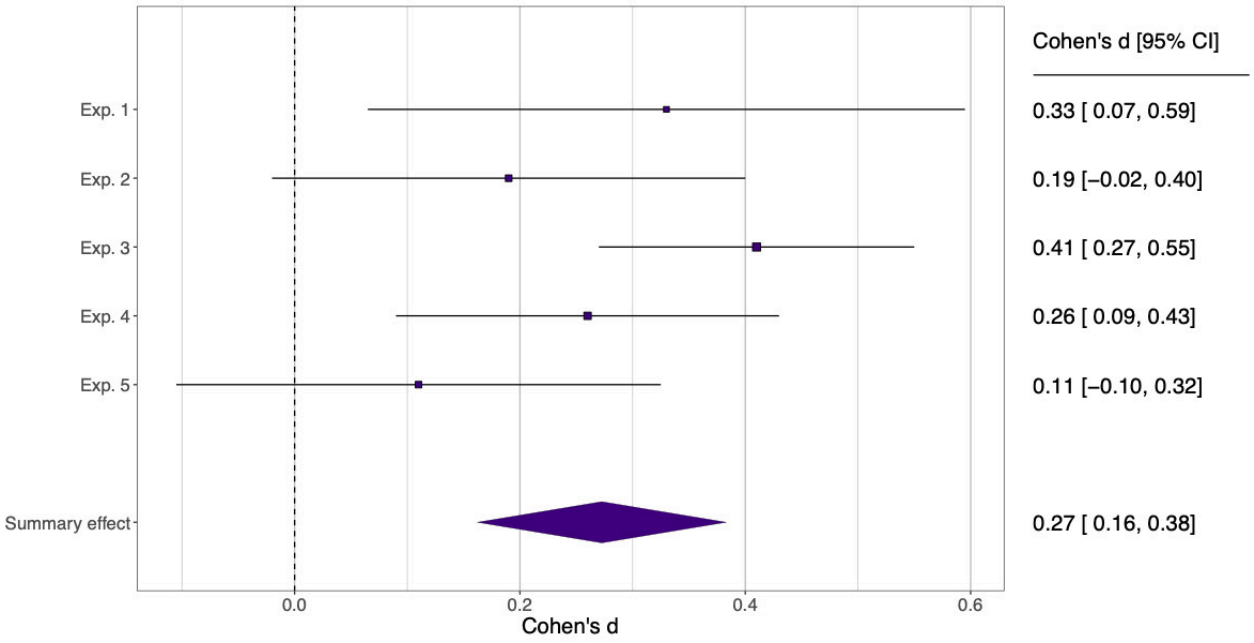

Figure 5. Average Effect Size of the Label Effect Across All Experiments

tween faces labelled as artificial or natural (Experiment 3). However, when participants only judged one type of face (i.e., either labelled as artificial or natural), the difference in trustworthiness judgments was eliminated (Experiment 5). The prime conclusion of the 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 judgment 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 results of Experiments 1-3 are difficult to reconcile with our initial hypothesis that the faces labelled as being artificial are represented as an outgroup, which results in these faces to be rated as less trustworthy (Balas & Pacella, 2017). At the same time, Experiment 4 further confirmed that the label effect is originated at the early stages of impression formation without the 'behavior' of the face coming at play. Finally, Experiment 5 suggests that this impression formation depends on the presence of different category of faces, which may indicate that some comparison or contrast between these categories is needed for the label effect to occur. One possible account that may fit these different findings is that the label 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.

Although we tested several boundary conditions of the label effect, the present study was restricted to the use of explicit ratings. Such ratings are assumed to reflect higher-order processes of deliberate reasoning, which may be affected by task demands. In contrast, attitudes may also result from automatic processes that occur spontaneously and outside of people's awareness or control (Bargh & Williams, 2006; Moors & De Houwer, 2006). Such 'implicit' attitudes are typically inferred from people's performance on response latency measures, such as the Implicit Association Test (IAT; Greenwald et al., 1998) or sequential priming tasks (Fazio et al., 1995; Wittenbrink et al., 1997). Importantly, explicit and implicit attitudes often do converge. For instance, Van Dessel et al. (2020) demonstrated that ex-

plicit attitudes towards social groups could be changes, but not implicit attitudes. Previous research already started to explore implicit attitudes in the context of social robotics (e.g., Diana et al., 2022; Erel et al., 2019) and question remains whether the label effect can be generalized when using more implicit measures.

A common limitation in the face perception literature is that stimulus materials are often relatively restricted with respect to the ethnicity of the faces used (Cook & Over, 2021). In the present study, our stimulus sets encompassed several ethnicities (Experiments 1-3 and 5), which is more in line with the contemporary multi-ethnic reality (e.g., Hong & Cheon, 2017). However, for the exception of Experiment 5, our experiments were not fine-tuned to counterbalance the proportions of different face ethnicities in the stimulus sets. The absence of counterbalancing introduced some biases in the stimulus sets where stimuli were selected on the basis of attractiveness and trustworthiness ratings (Experiments 1-3). Such ratings may reflect stereotype-driven biases (Lewis, 2011; Schmid et al., 2022; Sutherland et al., 2015) that result into a disproportionate amount of faces of a particular ethnicity to fall into a specific level of the judgment dimension (e.g., more Balck faces than White faces in high attractiveness categories; see Lewis, 2011). Nevertheless, this limitation is unlikely to have been the main source of the label effect for several reasons. First, we replicated the label effect under different conditions of stimulus variability (unlike in Experiments 1, 2, and 5, each participant in Experiment 3 was exposed to a set composed of different faces). Second, the effect also emerged in Experiment 4 with a different stimulus conveying a richer set of social cues, despite of its lower diversity (i.e., French-speaking natives of European descent) and imbalanced face sex (although consistently so cross label conditions). Moreover, the stimuli in Experiment 4 were manipulated to be more homogeneous in perceived trustworthiness, thus preventing the disproportionate allocation of any ethnicity to specific levels of a judgment. Finally, although it is tempting to conclude that the introduction of the optimal counterbalancing of social cues in Experiment 5 may have contributed to reduce or eliminate the label effect, such an explanation would remain harder to conciliate with the effect found with a sample that was more homogeneous in terms of physical appearance and perceived trustworthiness in Experiment 4. Instead, we believe that a contrast effect and/or lower believability in the nature of the stimuli in the computer-generated condition, are more likely to have been the factors resulting in the absence of a label effect in Experiment 5.

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.

Contributions

Contributed to conception and design: BL, MO, RH, HA, LML, EH

Contributed to acquisition of data: MO, LML, EH

Contributed to analysis and interpretation of data: MO, BL, LML, EH

Drafted and/or revised the article: BL, MO, RH, HA

Approved the submitted version for publication: BL, MO, RH

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/

Acknowledgments

We thank Alicia Sahel and Yanis Sahel for their valuable help with the translation and verification of the instruction materials. We would also like to thank the SHP4 students from Utrecht University for all the discussions and points they made that eventually landed in the paper. Finally, we thank the editor and reviewers for their insightful comments and suggestions to improve the quality of the article.

Funding

Part of this research was supported by Human-AI Alliance at Utrecht University.

Submitted: June 10, 2022 PDT, Accepted: March 01, 2023 PDT

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

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256

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

Balas, B., Tupa, L., & Pacella, J. (2018). Measuring social variables in real and artificial faces. *Computers in Human Behavior*, *88*, 236–243. https://doi.org/10.1016/j.chb.2018.07.013

Bargh, J. A., & Williams, E. L. (2006). The automaticity of social life. *Current Directions in Psychological Science*, *15*(1), 1–4. https://doi.org/10.1111/j.0963-7214.2006.00395.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4*. arXiv. https://arxiv.org/abs/1406.5823

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

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*, *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*, *11*(4), 147470491301100. https://doi.org/10.1177/147470491301100410

Cook, R., & Over, H. (2021). Why is the literature on first impressions so focused on White faces? *Royal Society Open Science*, *8*(9), 211146. https://doi.org/10.1098/rsos.211146

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 233. https://doi.org/10.1038/s41467-017-02597-8

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

DeBruine, L. M. (2022). *WebmorphR: Reproducible Stimuli* (0.1.1) [R]. https://CRAN.R-project.org/package=webmorphR

Diana, F., Kawahara, M., Saccardi, I., Hortensius, R., Tanaka, A., & Kret, M. E. (2022). A cross-cultural comparison on implicit and explicit attitudes towards artificial agents. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00917-7

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

Erel, H., Shem Tov, T., Kessler, Y., & Zuckerman, O. (2019). Robots are Always Social: Robotic Movements are Automatically Interpreted as Social Cues. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312758

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. https://doi.org/10.1037/0022-3514.69.6.1013

Ferdenzi, C., Delplanque, S., Mehu-Blantar, I., Da Paz Cabral, K. M., Domingos Felicio, M., & Sander, D. (2015). The Geneva Faces and Voices (GEFAV) database. *Behavior Research Methods*, *47*(4), 1110–1121. https://doi.org/10.3758/s13428-014-0545-0

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

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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*, *69*, 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

Hong, Y., & Cheon, B. K. (2017). How does culture matter in the face of globalization? *Perspectives on Psychological Science*, *12*(5), 810–823. https://doi.org/10.1177/1745691617700496

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

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521. https://doi.org/10.1038/s42256-019-0113-5

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, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

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

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology*, *70*(1), 47–65. https://doi.org/10.1037/0022-3514.70.1.47

Kossmeier, M., Tran, U. S., & Voracek, M. (2020). *metaviz: Forest Plots, Funnel Plots, and Visual Funnel Plot Inference for Meta-Analysis* (R package version 0.3.1). https://CRAN.R-project.org/package=metaviz

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

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

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *Package "Emmeans", R package version 4.0-3.*

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

Lewis, M. B. (2011). Who is the fairest of them all? Race, attractiveness and skin color sexual dimorphism. *Personality and Individual Differences*, *50*(2), 159–162. https://doi.org/10.1016/j.paid.2010.09.018

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

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

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

Miwa, K., & Terai, H. (2012). Impact of two types of partner, perceived or actual, in human–human and human–agent interaction. *Computers in Human Behavior*, *28*(4), 1286–1297. https://doi.org/10.1016/j.chb.2012.02.012

Moors, A., & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

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

Ng, Y.-L. (2022). When communicative AIs are cooperative actors: A prisoner's dilemma experiment on human–communicative artificial intelligence cooperation. *Behaviour & Information Technology*, 1–11. https://doi.org/10.1080/0144929x.2022.2111273

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

Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online. *Frontiers in Psychology*, *11*, 568256. https://doi.org/10.3389/fpsyg.2020.568256

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

Schmid, I., Witkower, Z., Götz, F. M., & Stieger, S. (2022). Registered report: Social face evaluation: ethnicity-specific differences in the judgement of trustworthiness of faces and facial parts. *Scientific Reports*, *12*(1), 18311. https://doi.org/10.1038/s41598-022-22709-9

Schwarz, N., & Bles, H. (1992). Scandals and the public's trust in politicians: Assimilation and contrast effects. *Personality and Social Psychology Bulletin*, *18*(5), 574–579. https://doi.org/10.1177/0146167292185007

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

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology*, *106*(2), 186–208. https://doi.org/10.1111/bjop.12085

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.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (2022). On the realness of people who do not exist: The social processing of artificial faces. *iScience*, *25*(12), 105441. https://doi.org/10.1016/j.isci.2022.105441

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *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

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, *72*(2), 262–274. https://doi.org/10.1037/0022-3514.72.2.262

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

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/73066-are-natural-faces-merely-labelled-as-artificial-trusted-less/attachment/152448.docx?auth_token=xEttFySaddrfCLr4Efo8