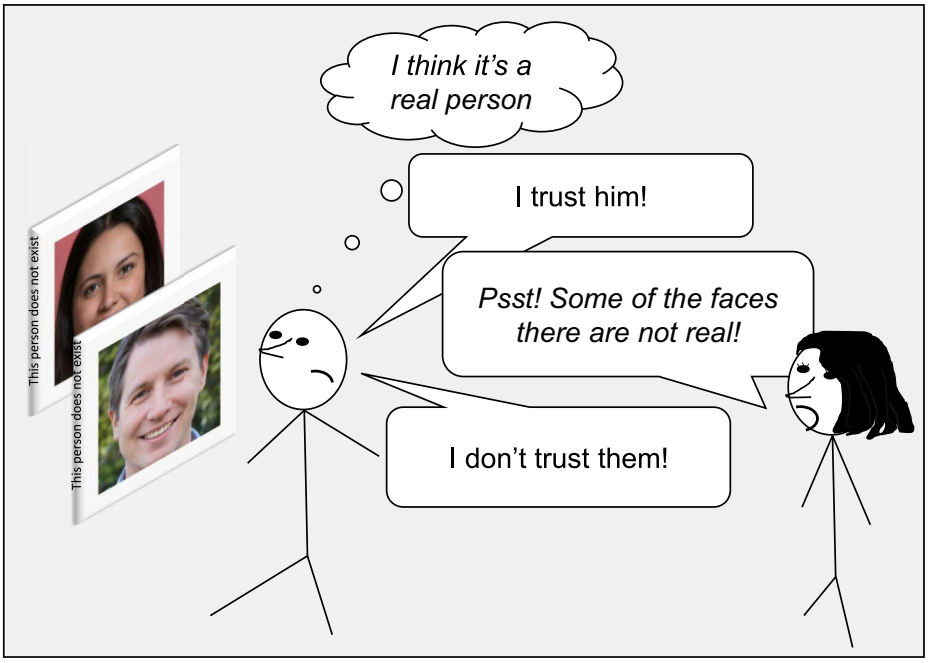
iScience

CellPress
OPEN ACCESS

Article

On the realness of people who do not exist: The social processing of artificial faces



Raffaele Tucciarelli, Neza Vehar, Shamil Chandaria, Manos Tsakiris

rtucciarelli@gmail.com

Highlights

GAN faces are more likely to be perceived as real faces than real faces

People are more likely to conform to faces that they had judged to be real

Informing people about the existence and presence of GAN faces erodes trust

Tucciarelli et al., iScience *25*, 105441
December 22, 2022
https://doi.org/10.1016/j.isci.2022.105441

iScience

CellPress
OPEN ACCESS

Article

On the realness of people who do not exist: The social processing of artificial faces

Raffaele Tucciarelli,[1,6,*] Neza Vehar,[1] Shamil Chandaria,[3,4] and Manos Tsakiris[1,2,5]

SUMMARY

Today more than ever, we are asked to evaluate the realness, truthfulness and trustworthiness of our social world. Here, we focus on how people evaluate realistic-looking faces of non-existing people generated by generative adversarial networks (GANs). GANs are increasingly used in marketing, journalism, social media, and political propaganda. In three studies, we investigated if and how participants can distinguish between GAN and REAL faces and the social consequences of their exposure to artificial faces. GAN faces were more likely to be perceived as *real* than REAL faces, a pattern partly explained by intrinsic stimulus characteristics. Moreover, participants' realness judgments influenced their behavior because they displayed increased social conformity toward faces perceived as *real*, independently of their actual realness. Lastly, knowledge about the presence of GAN faces eroded social trust. Our findings point to potentially far-reaching consequences for the pervasive use of GAN faces in a culture powered by images at unprecedented levels.

INTRODUCTION

More than ever before in human history, we are required to evaluate the realness, truthfulness and trustworthiness of our social world. From mainstream news to social media, from edited photos to deep fake videos, from humans to bots, and from alternative facts to fake news, we must judge the veracity of agents and the information they convey. In this work, we focus on people's ability to correctly judge the realness of artificially generated faces, that is, faces of people who do not exist.

Faces are among the most salient social stimuli we encounter in everyday life. We use people's facial features to form first impressions within a few milliseconds and infer personality traits (Palermo and Rhodes, 2007). We also use facial cues to judge people's attractiveness and trustworthiness (Todorov et al., 2015), to infer more complicated mental states and emotions (Hoffmann et al., 2010) and even to assess their health (Stephen et al., 2009). In addition to face-to-face encounters, the presence of digital faces is ubiquitous in modern visual culture, from traditional (e.g., television and magazines) and social media (e.g., Facebook and Twitter) to dating apps, from advertising to political marketing. In many cases, digital faces are as prevalent as faces in real-life encounters, and we regularly choose, edit and manipulate photos to communicate specific aspects of our personality. People are presumably not just interested in faces but in the minds behind those faces (Looser and Wheatley, 2010). Whenever we see a face, we automatically assume that the person in the photo *exists* (or existed at some point) and that this *real* person has a mind, thoughts, and emotions (Gray et al., 2007). However, with the emergence of the generative adversarial network (GAN) technology (Goodfellow et al., 2020; Karras et al., 2018), judging the veracity of faces has become even more challenging with potentially far-reaching societal consequences.

GANs are deep neural networks (i.e., artificial neural networks with multiple layers between the input and the output) that generate novel images which aim to be statistically indistinguishable from their training image dataset. GANs are particularly impressive at generating realistic novel faces trained on a large dataset of real faces. GANs are a machine learning technique in which two deep neural networks (DNNs) work in concert to train each other (see Figure 1). A *generative* network generates a fake face based on a random seed. The *discriminator* network is fed either the fake generated face or a real face (randomly drawn from a large training set of real faces) and has to decide if the face is fake. The generative model's task is to generate fake faces that will 'fool' the discriminator into classifying the face as real, hence the word *adversarial* in GAN. The discriminator classification error signal, i.e., the signal indicating if the discriminator is

[1]The Warburg Institute, School of Advanced Study, University of London, London WC1H 0AB, UK

[2]Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, UK

[3]Institute of Philosophy, School of Advanced Study, University of London, London, UK

[4]Centre for Psychedelic Research, Imperial College London, London, UK

[5]Centre for the Politics of Feelings, School of Advanced Study, University of London, London, UK

[6]Lead contact

*Correspondence: rtucciarelli@gmail.com

https://doi.org/10.1016/j.isci.2022.105441

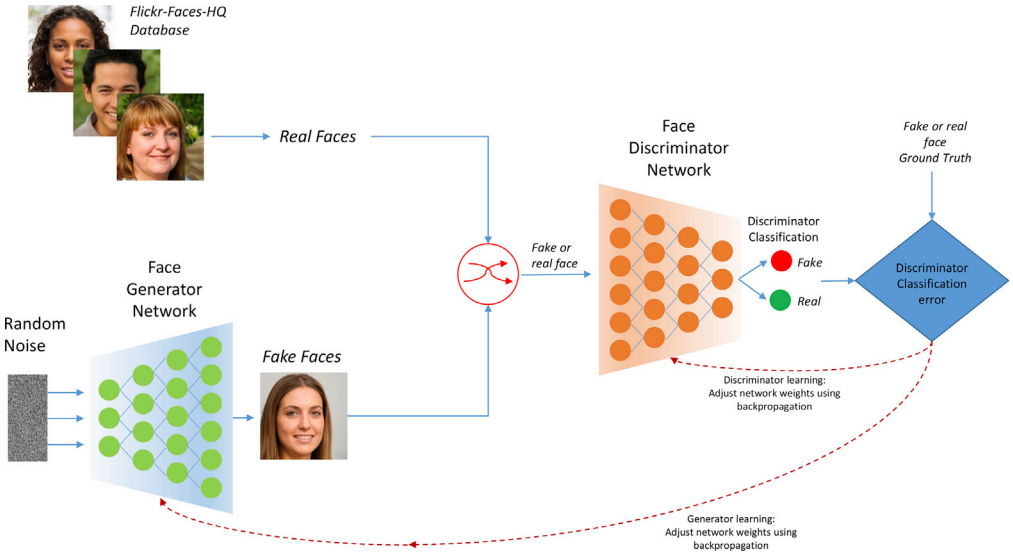


Figure 1. The figure illustrates how the face generator network is trained. See text for details.

mistaken in its classification of the face as real or fake, trains both the discriminator and generative networks by adjusting the network weights. This process is continued until both networks are proficient at their tasks and the generative network produces faces that are difficult to discriminate from real ones (also see sections 1 and 2.1 in the Data S1 file).

Over the last five years, there has been an impressive acceleration in GAN technology and, as a result, GAN faces look surprisingly realistic today (Beridze and Butcher, 2019). The people depicted in these new photos *do not exist* and never existed, and yet faces synthesized by GANs can look very much like the stimuli used to train the networks (Webster et al., 2021). This deep fake technology has been applied in various useful contexts (e.g., improving the quality of old photos, generating images for commercial websites, etc.), but the faces of non-existent people can and have been used with malicious intent, including in fake social media profiles (Rothman, 2018), that may influence social and political behavior. Images of artificial people are already ubiquitous in marketing, journalism, political propaganda (Vincent, 2020) and appear as "red herring" agents in intelligence wars (Parello-Plesner, 2018), where the intention is to influence, mislead or distract viewers (Satter, 2019). Alongside technical conventions and policy regulation for synthetic content, we need research that investigates how we process and behave toward GAN faces.

To the best of our knowledge, the experiments reported here go beyond previous research on fictional or computer-generated faces (Abraham, 2015; Balas and Tonsager, 2014; Green et al., 2008; Seyama and Nagayama, 2009). We investigated the extent to which people can identify state-of-the-art, hyper-realistic GAN generated artificial faces for what they are (i.e., not real) (Study 1, see preregistration at https://osf.io/5hswy). We then studied the social consequences of being unwittingly exposed to artificial faces, especially when they are perceived as real (Study 2, see preregistration https://osf.io/hae8q). We were particularly interested in people's informational conformity behavior (Toelch and Dolan, 2015), that is, people's tendency to copy sources that are deemed of higher informational value (Castelli et al., 2001; Paladino et al., 2010; Vaes et al., 2003). Is there a potential bias to rely on artificial faces as a source of information under conditions of uncertainty? The emergence of deep fake technology affords us a unique opportunity to probe potential differences in conforming to real and artificially generated social stimuli. For this purpose, we used a social conformity task to investigate how much and why people tend to conform to GAN over REAL faces, as a function of the perceived realness. Finally, we investigated whether by explicitly informing people about the presence of GAN faces and thus leading them to question social trust, we could influence the way in which they evaluated the information provided by GAN or REAL agents. To that end, in Study 3 (see preregistration at https://osf.io/x85pr), we investigated people's social conformity when they do or do not possess knowledge about the presence of artificial faces, in an attempt to understand whether participants conform more to non-existing than to real people, even when they know that some of the faces they see are artificial.

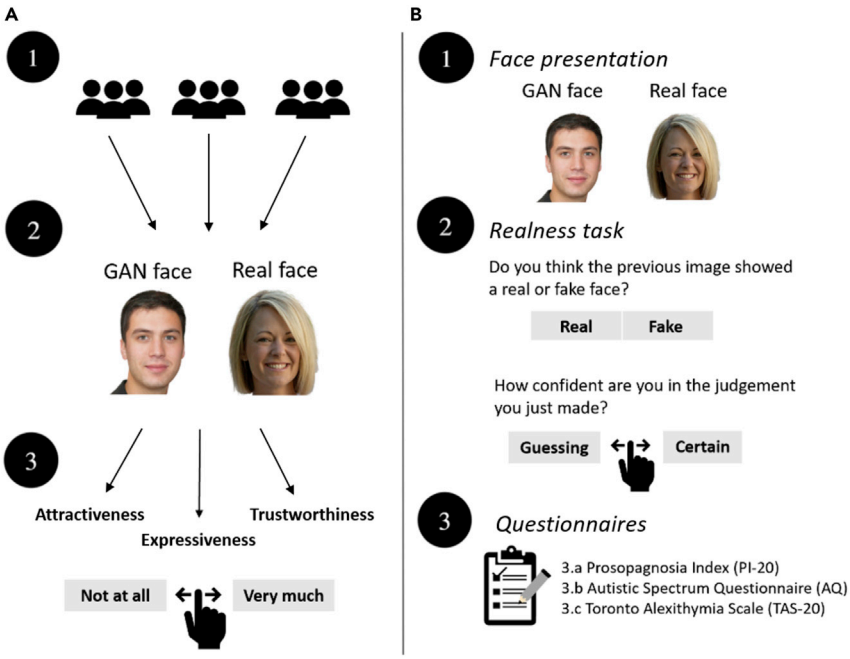


Figure 2. Steps in design of Study 1

(A) Before Study 1, we first conducted two laboratory pilot Stimuli Selection studies on independent groups of participants. In the first one, participants'($Npilot_1 = 4$) responses were used to choose the 50 out of 100 GAN stimuli rated highest on perceived realness. In the second study, responses from a new group of participants (Npilot2 = 4), were used to choose the 50 out of 100 REAL faces which rated highest on perceived realness (see sections 2.2 and 8 of the Data S1 file). We then used these faces in the three Stimuli validation studies: We recruited three different groups of naïve participants ($N_A = 31$, $N_E = 30$, $N_T = 30$), and exposed them to 100 faces (they did not know that half of them were GAN), each displayed for 3 s. Each group judged the Attractiveness, Expressiveness or Trustworthiness (e.g., from 'Not at all' to 'Very Attractive'), respectively (see section 2.4 of the Data S1 file).

(B) We then performed Study 1. (1) Participants were exposed to one face at a time, presented for 3 s. (2) They had to judge whether the face was 'Real' or 'Fake' and how confident they were in their answer (from 'Guessing' to 'Certain'), within 20 s. (3) We also obtained participant's scores on prosopagnosic, autistic and alexithymic traits, to investigate if and how such individual differences contribute to a different perception of Realness (for further details see STAR methods and section 2.7 of the Data S1 file).

RESULTS

Study 1: GAN faces are perceived as *real* faces

Participants were presented with GAN or REAL faces. They were instructed to judge whether these were *real* or *fake*, and then rate their confidence in their judgment (see Figure 2A). Following the main task, participants completed three questionnaires which assessed propensity to autism, prosopagnosia, and alexithymia respectively (Figure 2B).

Throughout our article, we use the definition of *real* (or REAL) to denote faces of people that actually exist (or existed), and *fake* to describe faces belonging to a non-existent person (i.e., GAN faces). Note that the distinction is often not obvious: A real face might look fake and a fake face might look real. In this sense, we expected that the categorical border between these two categories would not be clearly defined but rather ambiguous, and that *Realness* (i.e., the perception of reality from an image) would vary gradually across a continuum. Our aim was to investigate people's ability to recognize the realness of state-of-the-art GAN faces similar to those likely to be used for commercial, social, political or marketing purposes, rather than flawed GAN faces. Because this was central to our research question, the selection procedure for face images (see sections 2.1, 2.2 and 8 of the Data S1 file to see how GAN and REAL stimuli were obtained and for exclusion criteria) was conducted in a principled way instead of relying on random selection. As part of the selection procedure, we also obtained ratings for the levels of *Attractiveness*, *Expressiveness* and *Trustworthiness* of each face from independent raters (see Figures 2B, and S1 of the Data S1 file). These traits were specifically selected, because we wanted to account for possible stimulus characteristics that

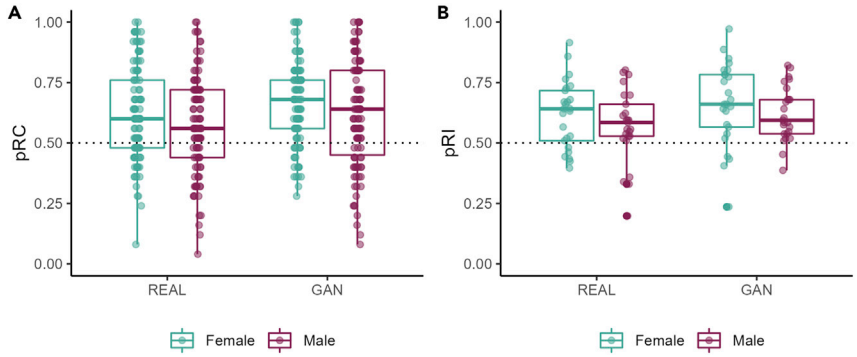


Figure 3. Study 1 – Descriptive plots

(A) By-participant boxplots. For each participant (i.e., each dot), we estimated the "Realness" (pRC) associated with each StimulusType and Stimulus Gender by dividing the number of "real" responses by the total number of faces within each category (N = 50, as we collapsed across Stimulus Gender).

(B) By-Image boxplots. For each image (i.e., each dot), we estimated the probability of being judged as real by dividing the number of participants that said "real" to that image by the number of participants. The dotted line is the chance level and indicates that a participant chose the faces within a category as "real" and "fake" equally as often.

might contribute to Realness judgments. Such traits have been observed to be implicated in the perception of computer-generated faces (Balas and Pacella, 2017; Balas et al., 2018). We tried to match and control for a number of face characteristics (e.g., age, smile, orientation; see section 2.3 of the Data S1 file and Table S1 of the Supplemental Information). The analyses are also described in the paragraph Analyses for the Realness task in STAR methods.

We first tested whether the two stimulus types were considered as *real* or *fake* relative to chance level. For each participant, we computed the proportion of *real* responses for each condition (pRC) by counting the number of *real* responses divided by the total number of images within a *StimulusType* (GAN| REAL) and *Stimulus Gender* (N = 25 repetitions for each combination of *Type* and *Gender*). The distributions of pRCs across conditions are shown in Figure 3A. The proportion of *real* responses for the Female/REAL Group was not normally distributed, as indicated by the *Shapiro-Wilk normality* test ($W_{Female/REAL}$ = 0.97, p = 0.016; $W_{Female/GAN}$ = 0.98, p = 0.09; $W_{Male/REAL}$ = 0.99, p = 0.53; $W_{Male/GAN}$ = 0.98, p = 0.07). We therefore used the nonparametric *Wilcoxon signed rank one-sample test* against 0.5 which resulted positive and significant for all distributions ($V_{Female/REAL}$ = 4516, p < 0.001, rank biserial correlation coefficient = 0.59, 95% CI = [0.43, 0.72]; $V_{Female/GAN}$ = 5163.5, p < 0.001, rank biserial correlation coefficient = 0.82, 95% CI = [0.74, 0.88]; $V_{Male/REAL}$ = 3998, p < 0.001, rank biserial correlation coefficient = 0.41, 95% CI = [0.21, 0.57]; $V_{Male/GAN}$ = 4429.5, p < 0.001, rank biserial correlation coefficient = 0.56, 95% CI = [0.39, 0.69]). We computed another index that captured the amount of realness associated to each image (pRI) by counting the number of participants that said *real* to each image divided by the number of participants. The distributions of pRIs across conditions are shown in Figure 3B. In this case, the proportion of *real* responses for all groups were normally distributed, as indicated by the *Shapiro-Wilk normality* test ($W_{Female/REAL}$ = 0.97, p = 0.574; $W_{Female/GAN}$ = 0.97, p = 0.758; $W_{Male/REAL}$ = 0.93, p = 0.112; $W_{Male/GAN}$ = 0.97, p = 0.732). The one-sample t-test against 0.5 was positive and significant for all groups ($T_{Female/REAL}$ = 4.23, p < 0.001, Cohen's d = 0.85, 95% CI = [0.39, 1.33]; $T_{Female/GAN}$ = 4.74, p < 0.001, Cohen's d = 0.95, 95% CI = [0.48, 1.45]; $T_{Male/REAL}$ = 2.34, p=0.028, Cohen's d = 0.47, 95% CI = [0.05, 0.90]; $T_{Male/GAN}$ = 5.19, p < 0.001, Cohen's d = 1.04, 95% CI = [0.55, 1.55]). These results suggest that the faces were generally considered to be *real* independently of type or stimulus gender.

Then, we investigated whether GAN faces were perceived differently than REAL faces. As reported in our pre-registration (https://osf.io/5hswy), based on the nature of the stimuli, we expected people to judge a REAL face to be *real* more often than a GAN face general linear mixed models (GLMM) logistic regression analysis with *StimulusType* (GAN|REAL), *Stimulus Gender* (Female|Male), *Attractiveness*, Expressiveness, Trustworthiness, and the individual differences captured by the three questionnaires on *alexithymia* (TAS-20), *autistic traits* (AQ) and *prosopagnosia* (PI) to predict participants' Judgment of the stimuli as *real* or *fake*. Participant *Age* and *Gender* were also added into the model. We included two random error components: Variation in the intercept because of participants (Participant ID) and images (Image ID). The

total explanatory power of the model was substantial (Conditional R^2 = 0.291), and the part related to the fixed effects alone (Marginal R^2) was 0.052. We observed that the *StimulusType* (Odd Ratio = 1.77, SE = 0.31, p=0.001, 95% CI = [1.25, 2.49]), and the *Stimulus Gender* (OR = 0.62, SE = 0.10, p=0.003, 95% CI = [0.46, 0.85]) were significantly predictive of participant responses, with GAN faces predicting *real* responses more than REAL faces and Female faces predicting *real* responses more than Male faces, respectively. The participant's *Gender* (OR = 1.47, SE = 0.29, p=0.050, 95% CI = [1.00, 2.15]) was also statistically significant, suggesting that Male participants tended to say *real*more often than Female participants. *Attractiveness* was also significant (OR = 0.67, SE = 0.06, p < 0.001, 95% CI = [0.55, 0.80]), suggesting that the more attractive a face was, the less it was considered *real* (Figure S2B). We also observed a significant interaction between *StimulusType* and *Age* (OR = 1.14, SE = 0.06, p=0.009, 95% CI = [1.03, 1.26]), suggesting that the tendency to say that GAN faces were more *real* than REAL faces increased with age (see Figures S2A and S5). All other predictors of the models were not statistically significant (see Table S2). Furthermore, GAN faces were associated with higher confidence levels than REAL faces judged; However, this effect was mostly explained by the fact that the GAN faces were judged to be more real than the REAL faces (see section 3.1.2. of the Data S1 file; Table S3 and Figure S3 of the Supplemental information).To evaluate the general accuracy of the participants, we also computed the *sensitivity* and *positive predictive value*(PPV); these are reported in section 5 of the Data S1 file. Also, none of the low-level features that we estimated (i.e., stimulus age, face orientation, smiling, flaws; see section 2.3 of the Data S1 file and Figure S1 of the Supplemental Information) could significantly explain any of the observed variance of the participants' responses.

In summary, Study 1 showed that GAN faces were judged to be more *real* than the REAL faces, even when controlling for the specified stimulus characteristics and participants' demographics, and even though GAN faces were rated as more attractive (see section 2.4.2. of the Data S1 file and Figure S1A of the Supplemental information). Generally, faces that were more attractive were considered to be less *real*. Our results contrast with previous studies that have investigated how people judge real and artificial faces (Balas and Tonsager, 2014). Past research has typically shown that artificial faces are easily categorized as fake, unless some important factors are disrupted (e.g., contrast). However, over the last decade, GAN technology has advanced considerably, and the GAN faces we used represent a state-of-the-art that did not exist a few years ago. The increased likelihood of classifying a GAN face as real showcases the performative power of current deep fake technology. As our results suggest, GAN stimuli have intrinsic features that lead them to be perceived as more real, but also that individual differences, such as age and gender, can also influence people's judgments.

Study 2: Judgment of realness, but not stimulus type, increases conformity

If GAN faces are more likely to be perceived as *real*, what, if any, are the social consequences of processing and interacting with such faces? To answer this question, Study 2 investigated how people socially conform to GAN and REAL faces. Previous studies showed that conformity (i.e., the over-proportional bias to match or copy a source that is deemed to have high informational value) can be experimentally manipulated using normative influences, elicited by social expectations or rules (Andrighetto et al., 2018; Castelli and Zecchini, 2005; Demoulin et al., 2004; Toelch and Dolan, 2015; Vaes et al., 2011). For example, conformity increases if the counterpart involved in the experiment is perceived to be part of the in-group (Vaes et al., 2003) or if the counterpart is relevant to one's self (Castelli et al., 2003). We hypothesized that people would conform more to faces that they perceive as *real*. Thus, they should display more conformity toward GAN than REAL faces, because these were generally rated as more *real*.

To test this hypothesis, the same participants who took part in Study 1 were invited to participate in Study 2 three months later. N = 55 participants took part in a modified version of the number of letter estimation task (Castelli et al., 2001). This task requires participants to estimate how many letters are displayed on the screen under time constraints. It also provides them with an additional source of information, i.e., an estimate of the number of letters suggested by another person, typically represented by a picture of a face. The use of this additional information source in completing the task provides an indication of conformity, because it quantifies the participant's level of trust in the face. This experimental setup aids in eliciting social conformity behavior because restricted cognitive resources (Pendry and Macrae, 1994) allow implicit attitudes to influence a participant's explicit judgments about the faces because they judge certain faces to be more informative and trustworthy sources than others (Castelli et al., 2003; van Cappellen et al., 2011).

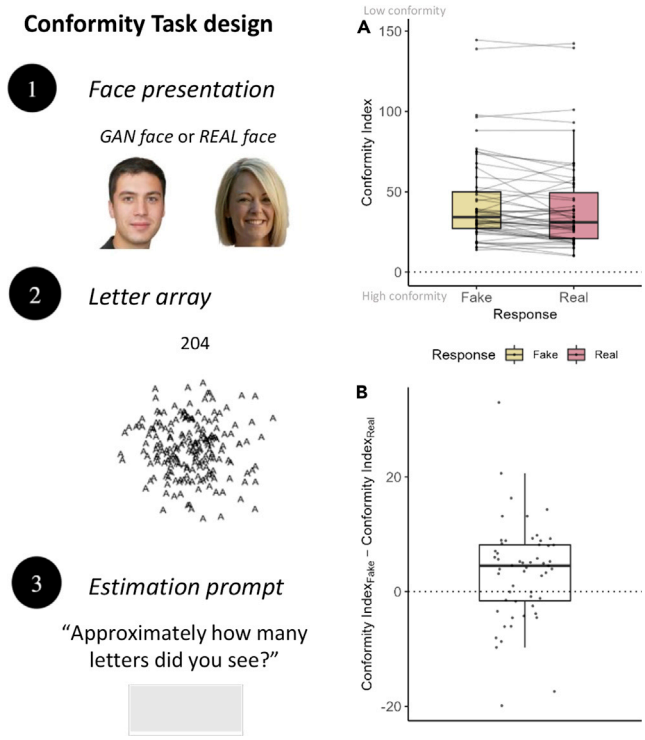


Figure 4. Study 2 – Conformity analysis

Design: (1) Sixty-four participants who previously participated in Study 1, also participated in the second experiment, but only those 54 participants that passed the attention checks were used in the analysis. They were exposed to the same 100 faces as before, followed by an array of letters 'A' (always showing 200 letters for 4 s) in different densities. (2) Above the letters, they saw a number which was an estimate of the previously seen face. Following this, they had to estimate how many letters were *on the screen within 10 s* (3) Descriptive plots: (A) For each participant, we first estimated the Conformity Index associated with "fake" and "real" responses and then averaged the Conformity Index across responses to obtain an estimate of the mean Conformity Index. Note that each pair of dots in Panel A represents a different participant. More specifically, they represent the averaged Conformity Index reported for the faces judged as fake and the averaged CI for the faces judged as real.

(B) For each participant, we then subtracted the mean Conformity Index associated with real responses from the Conformity Index with for fake responses. A value of zero (dotted line) indicates that there was on average no difference between the mean Conformity Indices. On average, the Conformity Index for images that participants judged to be fake was higher than for the ones judged to be real, suggesting that people were conforming more to faces that they judged to be real. Note that the analysis reported in panels (A and B) was exploratory (see the paragraph "exploratory analysis" for Study 2below).

On each trial, participants saw each of the 100 faces used in Study 1 for 3 s, followed by a dense cloud of letters (all *As*). Participants were asked to estimate the number of letters shown (Figure 4, left panel). The number of letters was always the same (N = 200), but the spatial distribution varied on each trial to give the impression that the number of *As* varied (Anobile et al., 2014). Crucially, at the top of the letter array there was a number representing the estimate provided by the face that had been presented earlier. As in past social conformity studies, participants could use this information to provide their answer. We informed participants beforehand that "*real faces tried very hard to provide a correct estimate*", and therefore the number could be informative, but the number provided by GAN faces was randomly generated by an algorithm, and therefore was not necessarily informative (see methods summary, *Conformity task* in STAR methods for further details). To incentivize participants to use all available information, we rewarded their accuracy with a bonus on a randomly selected trial. Conformity was measured as the absolute difference between the response and the estimate provided by the face, hereafter the *Conformity Index*. Thus, lower Conformity Index values indicate higher conformity.

As stated in our pre-registration (https://osf.io/hae8q), we first computed each participant's *Conformity Index* for the GAN and REAL faces (collapsing across *Stimulus Gender*) and then compared the two groups

using a paired test. Because the pair differences were not normally distributed (Shapiro-wilk test: W = 0.8, p < 0.001), we ran a *Wilcoxon signed rank test*, that resulted non-significant (V = 795, p = 0.837, rank biserial correlation coefficient = 0.03, 95% CI = [-0.26, 0.32]). Therefore, we could not reject the null hypothesis of no differences between the two group measures.

Exploratory analysis

With our planned analyses, we did not observe a significant difference between the *Conformity Index* elicited when participants saw GAN versus REAL faces. However, in our pre-registered analysis, we did not take into account people's responses on whether a face was perceived as *real* or *fake*. Therefore, we conducted an exploratory analysis to test whether there was a tendency to conform more to those faces that were judged as *real* (Figure 4, right panel for descriptive plots). We ran a GLMM regression analysis with a *negative binomial* distribution (*hurdle* regression analysis) to account for the fact that our dependent variable (*Conformity Index* scores) was bounded at zero. The advantage of the hurdle regression analysis is that it essentially allows one to run two models at once: one component (the *conditional* model) processes only the values larger than zero; the other component (the *zero-inflation* model) deals with the zero values (see analyses for the conformity task (study 2 and 3) in STAR methods for details). In our design, a zero value indicated maximum conformity. Therefore, we assumed that the zero-inflation model would inform us whether a participant conformed (i.e., *Conformity Index* = 0) or not (i.e., *Conformity Index*> 0). On the other hand, the conditional model quantifies the amount of conformity from high (low *Conformity Index* values) to low (high *Conformity Index* values). The model included the *StimulusType* (REAL|GAN), the *Stimulus Gender* (Female|Male), the *Judgment* (fake|real), and the interaction between *Judgment* and *StimulusType*, and between *Judgment* and *Stimulus Gender*, the *Gender* and *Age* of the participants as fixed-effects. We also included the *Attractiveness*, *Expressiveness and Trustworthiness* of the face as covariates. The *Participant* and *Image IDs* were used as random-effect intercepts. For the conditional model, the *Judgment* was significant (Incident Rate Ratio = 0.92, SE = 0.02, p=0.001, 95% CI = [0.88, 0.97]) suggesting smaller *Conformity Index* values (i.e., higher conformity) to faces that were judged to be *real* than to those judged to be *fake*. The *participant Gender* was also significant (IRR = 1.37, SE = 0.22, p=0.047, 95% CI = [1.00, 1.87]), suggesting that Male participants tended to have higher *Conformity Index* values (i.e., less conformity) than Female participants. The zero-inflated component also confirmed that more *Conformity Index* values equal to zero (i.e., maximal conformity) were given to faces that were judged to be *real* to those considered *fake* (Odd Ratio = 1.33, SE = 0.18, p=0.040, 95% CI = [1.01, 1.75]). *Attractiveness* was also significant (OR = 0.82, SE = 0.08, p=0.033, 95% CI = [0.69, 0.98]), indicating that participants tended to conform less to faces rated as more attractive. The *Stimulus Type*,*Stimulus Gender* and their interactions were not significant in either of the model components (see Table S4).

In sum, we observed that GAN faces were more likely to be judged as *real* than REAL faces (Study 1), and that participants (a subset of Study 1) were more likely to display greater social conformity to faces that they had previously judged as *real* (Study 2), independently of stimulus type.

Study 3: Informing people about the presence of GAN faces reduces conformity and attenuates trust

In Studies 1 and 2, participants were informed from the outset about the nature of GAN faces and their presence in our experiment. In that sense, the informational context within which they performed these experiments was one where they were encouraged to entertain doubts about the authenticity and potential trustworthiness of the perceived faces. This simulates the reality of our daily interactions with artificially-generated images. The fact that GAN faces were perceived to be more *real* showcases the recent technological advances in the application of Artificial Intelligence (AI) in image generation. As such technologies become more ubiquitous in social media, the very presence of artificial agents, bots, GAN images and deep fake videos may erode our trust in what we see and hear. In Study 3, we aimed to understand how knowledge about the nature and presence of such stimuli may impact social trust, by explicitly manipulating the informational context within which participants performed the experiment, either informing or not informing them at the outset about the presence and nature of GAN faces.

In Study 3, two new groups of naïve participants were recruited in a between-subjects experiment. In Unlike Studies 1 and 2, in Study 3, participants first performed the Conformity task, and then the Realness task. This allowed us to manipulate participants' knowledge about the types of faces we used. In the Knowledge group, participants were informed at the beginning of the experiment that some of the faces were

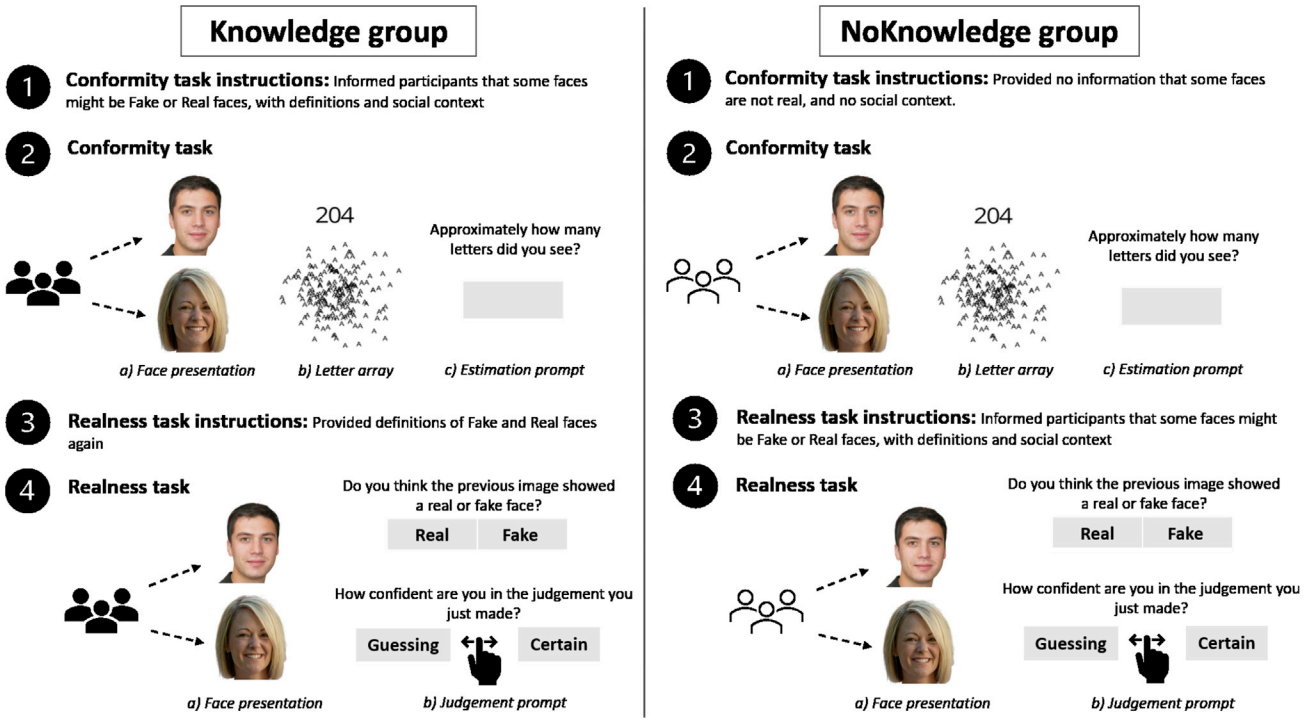


Figure 5. Design of Study 3:Participants were recruited for the Knowledge (N_K = 116) and the NoKnowledge (N_{NK} = 116) groups

(A) The Knowledge group was told that some of the faces were Fake, defined as: "Artificial faces that are generated by an algorithm and these images depict non-existing people", whereas others were Real, "Genuine, unaltered faces of real and existing people". This explanation was accompanied by comments on concurrent social context "Such technology of generating artificial faces of non-existing people can be used in various useful contexts (e.g., improve the quality of old photos, generate new model images for commercial websites, etc.), but is also being used with malicious intentions, such as generate fake social media profiles that could influence social and political behavior. It is therefore important and timely to investigate how we process such faces". The NoKnowledge group did not receive this information.

(B) Both groups performed the Conformity task, as described in methods – Study 2. When performing the conformity task, both groups were told that some people (represented by these *faces*) may have made more effort than others and that they should use any information available to them, to consider whether the person gave an informative response that they might want to take on board.

(C) Participants in the NoKnowledge group were told for the first time that some faces might be Real or Fake (with accompanying definitions and social context), whereas the Knowledge group received the same definitions again.

(D) Participants in both groups then performed the Realness task, as described in methods – Study 1.

artificially-generated (i.e., were GAN faces), whereas in the NoKnowledge group they received that information only before the Realness task (Figure 5). In that way, we were able to quantify people's social conformity to the same face stimuli under two different knowledge contexts and could assess whether knowledge about the presence of artificial agents impacts trust and conformity.

To compare whether informational context modulated the level of conformity, we first ran an independent Mann-Whitney U Test between the two groups, because the Conformity Index scores for both groups were not normally distributed, as indicated by the *Shapiro-Wilk normality* test ($W_{Knowledge}$ = 0.87, $p < 0.001$; $W_{NoKnowledge}$ = 0.77, $p < 0.001$). We observed that that the *Conformity Index* was on average larger (i.e., indicated less conformity) for the Knowledge than for the NoKnowledge group (W = 7825, p=0.016, rank biserial correlation = 0.18, 95% CI = [0.04, 0.32]) suggesting that knowing that GAN faces were present reduced trust in general (Figure 6). Then, as in Study 1, and as stated in the pre-registration (https://osf.io/x85pr), we ran general linear mixed models (GLMM) with a *binomial* distribution to predict judgment and a GLMM with a *Gaussian* distribution to predict confidence (see Tables S8–S13 for the results; See also Table S17 for a comparison across studies). Furthermore, as in Study 2, and as stated in the pre-registration (https://osf.io/x85pr), we ran GLMMs with a negative *binomial* distribution (*hurdle* regression analysis) to predict *Conformity Index* scores. The fixed-effects predictors were *StimulusType* (REAL|GAN), *Stimulus Gender* (Male|Female), the *Judgment* (Fake|Real), and their interactions with *Group* (Knowledge|NoKnowledge), the *Attractiveness*, the *Gender* (Female|Male) and *Age* of the participants. The *Participant*

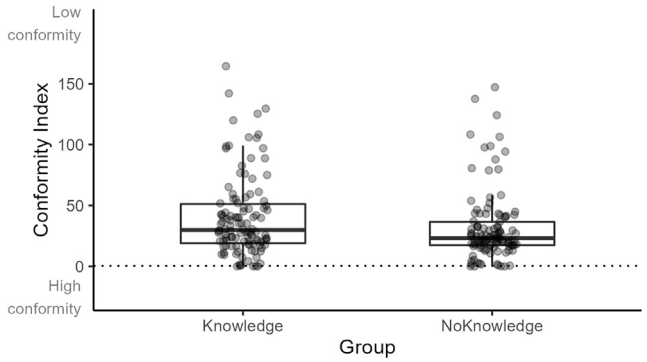


Figure 6. Conformity results of Study 3

Conformity Index scores for the two groups. The index was averaged across StimulusType and Stimulus Gender, each dot representing a participant. Smaller values indicate higher conformity, with zero indicating maximal conformity (i.e., the participant just copied the choice of the face); Higher values indicate lower conformity.

and *Image IDs* were used as random-effect intercepts. The same model definition was used for the zero-inflated component. For the conditional model, none of the parameters were significant. For the zero-inflated component, *Judgment* was significant (Odd Ratio = 1.25, SE = 0.10, p=0.005, 95% CI = [1.07, 1.45]) suggesting more *Conformity Index* values equal to zero (i.e., maximal conformity) to faces that were judged to be *real* than to those judged to be *fake* (Figure 7), thus confirming the results that we also observed in Study 2. We further observed a significant interaction between *StimulusType* and *Group* (OR = 0.79, SE = 0.08, p=0.025, 95% CI = [0.65, 0.97]), and a significant interaction between *Judgment and Group* (OR = 0.80, SE = 0.09, p=0.034, 95% CI = [0.65, 0.98]). To further characterize the interactions, we ran two GLMMs to fit the data of the two groups separately. This analysis revealed that, for the zero-component, *Judgment* of realness significantly predicted the Conformity Index for the *Knowledge* group (OR = 1.25, SE = 0.10, p=0.004, 95% CI = [1.07, 1.46]), but not for the *NoKnowledge* group (OR = 0.99, SE = 0.07, p = 0.933, 95% CI = [0.86, 1.15]), suggesting that the conformity toward faces believed to be *real* increased only when participants knew that artificial faces were present (note that also in Study 2 participants were aware of the presence of artificial faces). Furthermore, the zero-inflated *Age* predictor was significant in both groups (Knowledge: OR = 2.01, SE = 0.53, p=0.008, 95% CI = [1.20, 3.37]; NoKnowledge: OR = 1.79, SE = 0.51, p=0.043, 95% CI = [1.02, 3.13]), suggesting that the number of zeros, and consequently conformity, increased with age in both groups. Finally, for the NoKnowledge group only, the zero-inflated *Trustworthiness* predictor was significant (Knowledge: OR = 0.96, SE = 0.06, p = 0.462, 95% CI = [0.85, 1.08]; NoKnowledge: OR = 1.15, SE = 0.07, p=0.026, 95% CI = [1.02, 1.29]), suggesting that the number of zeros, and

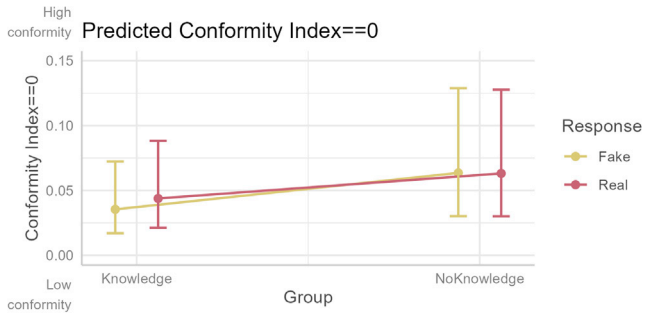


Figure 7. Predicted zero-inflation probabilities of Conformity Index for Study 3

We ran a GLMM hurdle regression analysis to explain the Conformity Index. Because this is the zero-inflated component of the model, higher values of probability associated with the Conformity Index would suggest a higher number of zeros (i.e., maximum conformity). Please note that this is why the labels "High conformity" and "Low conformity" are inverted with respect to Figures 4A and 6. We observed a significant interaction between Judgment and Group (OR = 0.80, SE = 0.09,p = 0.034, 95% CI = [0.65, 0.98]), suggesting that the Knowledge group only conformed more to the faces judged to be real rather than fake. See main text and Figure 8 for more details. Error bars indicate 95% confidence intervals.

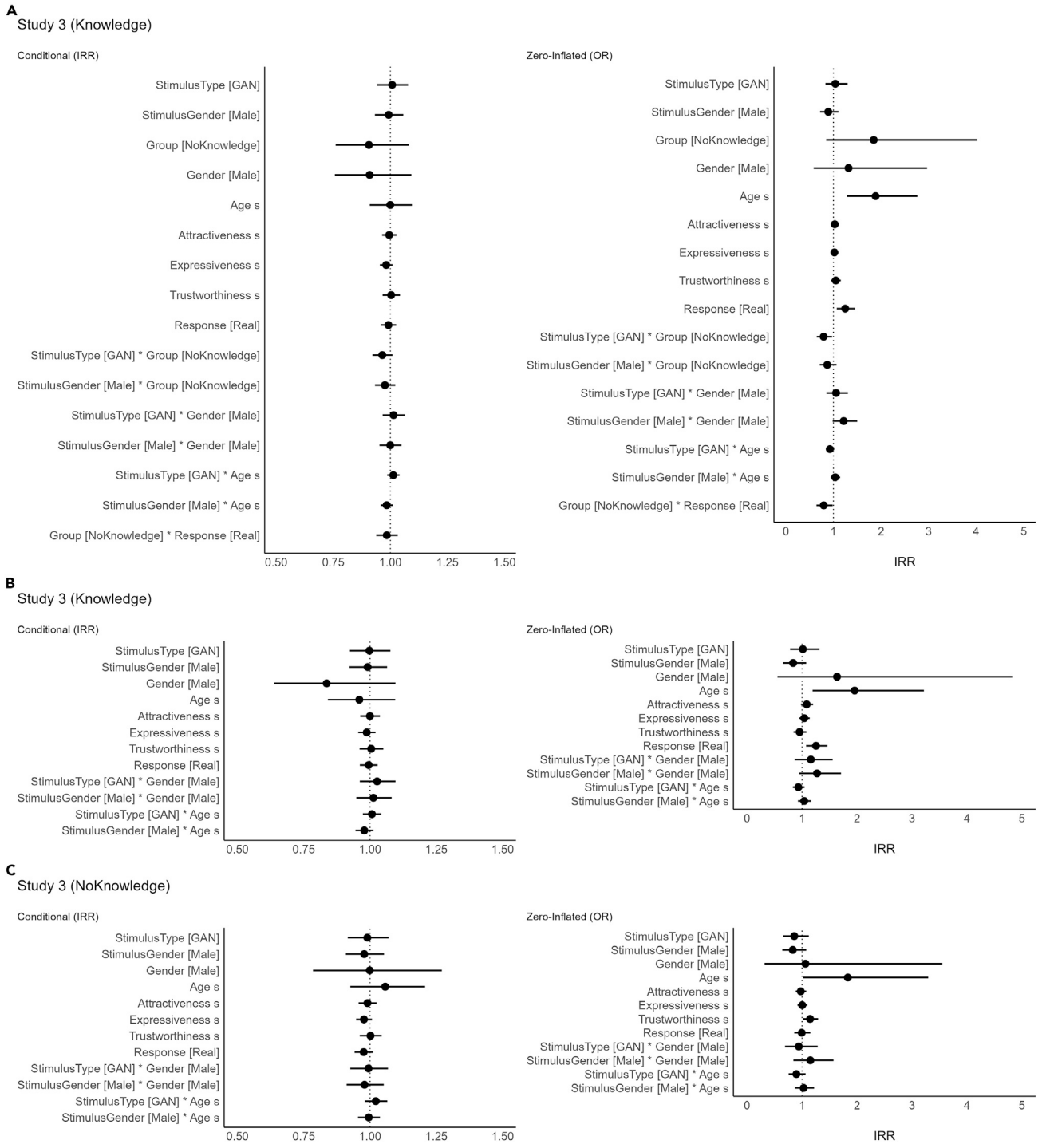


Figure 8. Estimated parameters of the hurdle analysis to predict the Conformity Index (Study 3)

(A) Estimated parameters for the conditional and zero-inflated components when the two groups (Knowledge and NoKnowledge) were used in the same model. For the conditional component, Incident Rate Ratios (IRR) are reported, therefore a value of 1 indicates no contribution in predicting the Conformity Index; a value significantly larger than 1 indicates that the parameter contributes to predicting higher values in the Conformity Index (i.e., lower conformity). For the zero-inflated component, we report the odds ratio (OR), where a value of 1 indicates no contribution in predicting the Conformity Index. Note that in this case the model predicts the probability of observing zeros (i.e., maximal conformity); a value significantly larger than 1 indicates that the parameter is contributes to predicting more zeros. For example, for the Group parameter we observed a value significantly larger than one. This suggests that more zeros

Figure 8. *Continued*
were observed for the NoKnowledge group than for the Knowledge group (the reference level), suggesting more conformity was associated with the NoKnowledge group.
(B and C) Panels (B and C) show the estimated parameters for the Knowledge and No Knowledge groups, respectively. Error bars indicate 95% confidence intervals.

consequently conformity, increased with the level of Trustworthiness of the faces, but only in the NoKnowledge group (Figure 8; see Table S5 for both groups; Table S6 for the Knowledge group; and Table S7 for the NoKnowledge group). For an overview on the *sensitivity* and the *positive predictive value* for this experiment, please refer to section 5 of the Data S1 file.

Study 3 therefore showed that the informational context within which people encounter artificial faces and prior knowledge of their presence impact trusting behavior in such social conformity settings. Overall, the NoKnowledge group displayed greater conformity than the Knowledge group. Thus, in the absence of knowledge about the presence of artificial agents, people displayed greater conformity. However, if they were informed and made aware of their potential presence, trust and conformity was diminished. Moreover, only the Knowledge group, as was the case in Study 2, tended to conform more to faces that they believed to be real, independently of stimulus type, whereas the NoKnowledge group showed greater conformity in general.

Study 4: Pre-selecting faces increases the probability of misclassification

To validate and strengthen the motivations for our stimuli pre-selection, we ran another online study, Study 4, with a new sample (N = 116). For this study, the procedure was similar to that used in Study 1 and the second part of Study 3, i.e., participants had to say whether a face was *real* or *fake*, but in this case, for each participant, we randomly sampled 20 GAN faces and 20 REAL faces from the full pool of 100 GAN and 100 REAL faces initially selected, meaning prior to the "stimuli selection procedure" (see section 2.1 of the Data S1 file). This means that the stimuli could contain flaws, obvious errors, filters, faces with strong make-up, etc. As for Study 1 and 3, we first computed the proportion of real for each stimulus type (see descriptive plots in Figure 9). We predicted that flawed stimuli would have be considered less *real* than the pre-selected GAN stimuli. We ran a new GLMM Logistic regression analysis to predict *Judgment* as a function of the *Preselection*(Flawed|Preselected), *StimulusType* (REAL|GAN), the *Stimulus Gender* (Male|Female), and the participant's *Gender* and *Age*. We also included the interaction between *Preselection* and *StimulusType*, between *Preselection* and *Stimulus Gender*, between *StimulusType* and *Age*, between *StimulusType* and *participants Gender*, between the *Stimulus Gender* and the *Age*, and between the *Stimulus Gender* and the participants Gender. As random-effects components, we included variation in the intercept because of participants (Participant ID) and images (Image ID). We observed that *Preselection* (OR = 1.64, SE = 0.32, p=0.012, 95% CI = [1.11, 2.40]), *StimulusType* (OR = 0.64, SE = 0.12, p=0.016, 95% CI = [0.44, 0.92]) and *Age* (OR = 0.82, SE = 0.08, p=0.043, 95% CI = [0.68, 0.99]) significantly predicted participants' responses. Furthermore, we observed that the interaction between *StimulusType* and *Preselection* (OR = 1.57, SE = 0.36, p=0.045, 95% CI = [1.01, 2.45]) and between *StimulusType* and *Age* (OR = 1.50, SE = 0.15, $p < 0.001$, 95% CI = [1.23, 1.81]) significantly predicted participants' responses. To better characterize the significant interactions, we ran two other GLMM Logistic regression analyses separately for the Preselected and Flawed faces using the same predictors as before, but excluding the *Preselection* predictor. For the Flawed faces, we observed a significant effect of *Age* (OR = 0.82, SE = 0.08, p=0.047, 95% CI = [0.68, 1.00]), and interaction between *Age* and *the StimulusType* (OR = 1.49, SE = 0.15, $p < 0.001$, 95% CI = [1.23, 1.81]). For the Preselected faces, we also observed a significant interaction between *Age* and *StimulusType* (OR = 0.75, SE = 0.07, p=0.004, 95% CI = [0.62, 0.91]); and *Attractiveness* (OR = 1.23, SE = 0.12, p=0.038, 95% CI = [1.01, 1.49]) was also a significant predictor of participants' responses. Note that we could use *Attractiveness* only for the Preselected faces because we had not collected this measure for the Flawed faces. The interaction between *Age* and *StimulusType* on both the Flawed and Preselected face models can be appreciated in Figure S5 (bottom plot). These plots seem to indicate that the probability of classifying GAN faces as *real* increased with *Age* (see also Figure S4). All estimated predictors are reported in Tables S14–S16). For an overview on the *sensitivity* and the *positive predictive value* for this experiment, please refer to section 5 of the Data S1 file.

These results suggest that, as expected, GAN faces are easily classified as *fake* when they contain obvious errors and flaws. Preselecting the stimuli in a principled way increased the probability that a GAN face would be selected as *real* and eliminated the difference between GAN and REAL faces.

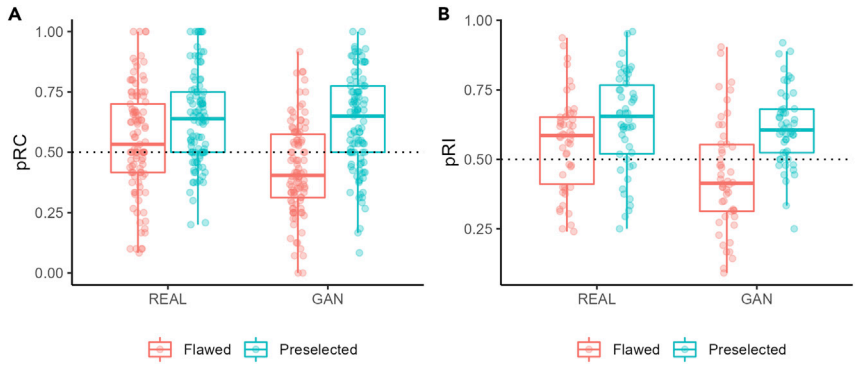


Figure 9. Study 4. – Descriptive plots

(A) By-participant boxplots. We used the same procedure as in Study 1 and 3 to compute the proportion of real responses for each subject (see Figure 3A). For each participant (i.e., each dot), we estimated the "Realness" (pRC) associated with each StimulusType and category (Flawed and Preselected) by dividing the number of "real" responses by the total number of faces within each category (N = 25). The Preselected stimuli were the faces selected during the stimuli selection procedure (see section 2.2. of the Data S1 file), whereas the Flawed stimuli were the ones that were discarded during the stimuli selection procedure.

(B) By-image boxplots. We used the same procedure as in Study 1 and 3 to compute the proportion of real responses for each subject (see Figure 3B).

Stimulus detection theory: Sensitivity and bias analysis across studies

To gain a better understanding of participants' ability to detecting REAL and GAN faces, we also ran sensitivity and bias analyses, but note that these were not pre-registered for any of the studies. A correct classification of a REAL face as *real* was considered a *hit*, and a correct classification of a GAN face as fake, a *correct rejection*. The sensitivity index (*d-prime* or *d'*) was computed as the difference between the normalized hit rate and the normalized false alarm rate (Macmillan and Creelman, 2004). This index gives an indication of how good participants were at detecting GAN and REAL faces. Positive values indicate good classification of the faces (REAL faces as *real* and GAN faces as *fake*); Negative values indicate bad classification of the faces (REAL faces as *fake* and GAN faces as *real*); Values equal to zero indicate that participants were not able to distinguish between the two stimulus types. The sensitivity index is an unbiased measure, meaning that it is not influenced by the decision criterion chosen by the participants while performing the task. The bias or decision criterion (c) index, which indicates a participant's tendency to be liberal (i.e., to say *real*) or conservative (i.e., to say *fake*) in their responses, was computed as the midpoint between the normalized hit and false alarm distributions multiplied by minus 1. Therefore, a value of zero indicates no bias or tendency to say *real* or *fake*. Values greater than zero indicate a tendency to be conservative, i.e., to say *fake* more often; Values smaller than zero indicate a tendency to be liberal, i.e., to say *real* more often. We examined the sensitivity (Figure 10) and bias indices (Figure 11) in Study 1, Study 3 and Study 4 where participants had to classify faces.

First, we examined whether the indices differed from zero using a one-sample t-test. In Study 1, we observed a very small but significant negative result for d-prime (Mean = −0.13, $t_{105.00}$ = −2.54, p=0.012, 95% CI = [−0.23, −0.03], Cohen's d: −0.25) and a medium significant negative result for the criterion used (Mean = −0.36, $t_{105.00}$ = −7.12, p < 0.001, 95% CI = [−0.46, −0.26], Cohen's d: −0.69), suggesting that participants tended to misclassify the faces (i.e., classified GAN faces as *real* and REAL faces as *fake*), but in general tended to say that faces were *real*.

In Study 2, for the Knowledge group, we observed a very small but significant positive result for d-prime in the Knowledge group (Mean = 0.15, $t_{114.00}$ = 2.70, p=0.008, 95% CI = [0.04, 0.26], Cohen's d: 0.25) and no significant difference for the NoKnowledge group (Mean = 0.05, $t_{115.00}$ = 0.96, p = 0.338, 95% CI = [-0.05, 0.14], Cohen's d: 0.09). However, we did not see a significant difference between the two groups when they were directly compared using an independent sample t-test (Mean1 = 0.15, Mean2 = 0.05, Difference = 0.11, $t_{222.10}$ = 1.45, p = 0.150, 95% CI = [−0.04, 0.25], Cohen's d: 0.19). Furthermore, for the Knowledge group, we observed a medium significant negative result for the criterion used in the Knowledge group (Mean = −0.26, $t_{114.00}$ = −7.23, p < 0.001, 95% CI = [-0.33, −0.19], Cohen's d: −0.67) and a medium

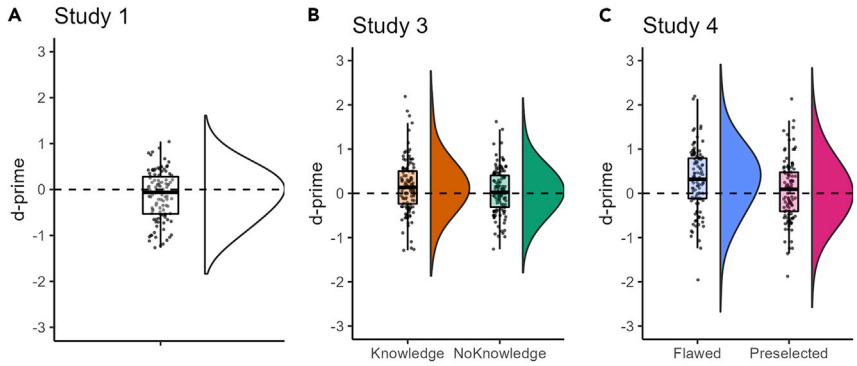


Figure 10. D-prime

(A–C) Distribution of d-prime index in the three Studies: (A) Study 1; (B) Study 3, with the Knowledge and NoKnowledge groups separated; (C) Study 4, with the sampling method separated. Each dot represents a participant in Study 1 and 3; In Study 4, each pair across sampling method represent a participant.

significant negative result for the NoKnowledge group (Mean = −0.28, $t_{115.00}$ = −7.47, $p < 0.001$, 95% CI = [-0.36, −0.21], Cohen's d: −0.69). We did not see a significant difference between the two groups for the criterion used (Mean1 = −0.26, Mean2 = −0.28, Difference = 0.02, $t_{228.71}$ = 0.38, p = 0.707, 95% CI = [−0.08, 0.12], Cohen's d: 0.05). These results seem to indicate that even if both groups used a similarly liberal criterion, the Knowledge group tended to classify the faces correctly (i.e., GAN faces as *fake* and REAL faces as *real*).

Finally, in Study 4, we observed a small significant positive result for d-prime in the Flawed condition (Mean = 0.30, $t_{104.00}$ = 4.39, $p < 0.001$, 95% CI = [0.17, 0.44], Cohen's d: 0.43) and no significant difference for the Preselected condition (Mean = 0.05, $t_{104.00}$ = 0.76, p = 0.450, 95% CI = [−0.08, 0.18], Cohen's d: 0.07); The difference between the two conditions was significant (Difference = 0.36, $t_{104.00}$ = 9.16, $p < 0.001$, 95% CI = [0.28, 0.44], Cohen's d: 0.89). Furthermore, we observed no significant difference for the criterion used in the Flawed condition (Mean = 0.01, $t_{104.00}$ = 0.33, p = 0.744, 95% CI = [−0.07, 0.09], Cohen's d: 0.03) and a large significant negative result for the Preselected condition (Mean = −0.35, $t_{104.00}$ = −8.81, $p < 0.001$, 95% CI = [−0.43, −0.27], Cohen's d: −0.86). We also observed a large significant difference when the two conditions were directly compared using a paired t-test (Difference = 0.36, $t_{104.00}$ = 9.16, $p < 0.001$, 95% CI = [0.28, 0.44], Cohen's d: 0.89). These results seem to indicate that faces that were randomly selected could be easily and correctly classified, most likely because faces with clear flaws were also sampled; on the other hand, the faces that were selected in a principled way (i.e.,

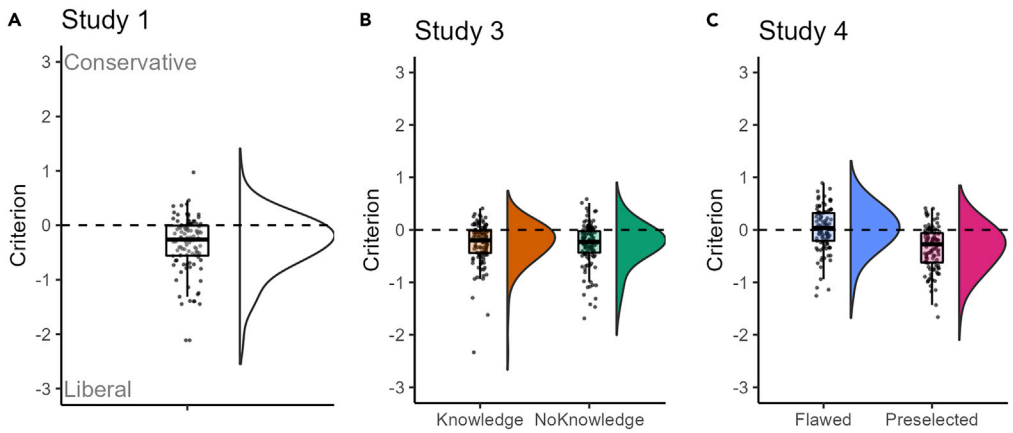


Figure 11. Bias

(A–C) Distribution of criterion (bias) index in the three Studies: (A) Study 1; (B) Study 3, with the Knowledge and NoKnowledge groups separated; (C) Study 4, with the sampling method separated. Each dot represents a participant in Study 1 and 3; In Study 4, each pair across Stimuli (Flawed and Preselected) sampling methods represents a participant.

same stimuli used in Study 1, 2, and 3) could not be discriminated. Of interest, the two conditions also differed for the criterion used, with participants being unbiased in the Flawed condition, suggesting that it was easier to discriminate between GAN and REAL faces, and participants exhibited a liberal bias in the Preselected condition, as observed in the previous studies.

DISCUSSION

Across three online studies, we investigated how people perceive and process artificial faces and the ensuing social consequences. GAN faces were more likely to be perceived as *real* than actual REAL faces, a finding that is consistent with a recent study using similar stimuli to ours (Nightingale and Farid, 2022). Moreover, people's perception of the realness of GAN faces made them more likely to conform, indicating higher trust, to faces that they had judged to be *real*, rather than to REAL faces per se. Lastly, informing people about the existence and presence of GAN faces erodes trust, yet people still conform more to faces they judge to be *real*, rather than to REAL faces per se. Our results offer novel insights into *how* such images are perceived as real and why, as well as on how their social use may influence behavior.

What makes GAN faces more real than REAL faces? Some of the observed variance in judgments of realness could be explained by the intrinsic characteristics of GAN faces per se. For example, perceived attractiveness negatively predicted realness judgment, because faces that were rated as less attractive were also rated as *more real*. Less attractive faces might be considered more *typical*. According to the norm-based face space theory (Valentine, 1991), the typical face has a special status and it is used as a reference to which all faces are evaluated. Therefore, these stimuli would look more real because they are more similar to mental templates that people have built from instances of faces seen in everyday life. Although in general typical faces are assumed to be more attractive (Rhodes, 2006), it was recently shown that attractiveness ratings decreased for faces closer to the typical face (Sofer et al., 2015). However, attractiveness could only explain some of the variance in judgment; stimulus type alone (GAN vs REAL) could still explain judgment when attractiveness was held constant across our two stimulus types. Although partly explained by stimulus features, the perceived realness of faces is also in the mind of the beholder insofar the participants' social cognition and face processing traits, included as control variables in our analysis, also influenced performance. For a more in-depth discussion on the role of individual differences, refer to section 4.1 of the Data S1 file.

Nightingale and Farid (2022) recently showed that GAN faces are indistinguishable from real faces in terms of their realness. In our studies, we were mostly interested in investigating the characteristics that lead to a face being perceived as real and, in particular, whether GAN faces were perceived to be more real than REAL faces. Another interesting question was about participants' performance in classifying the two stimulus types across studies and conditions. To answer this question, we used signal detection theory (Macmillan and Creelman, 2004). Across studies, participants either misclassified, i.e., d-prime was on average significantly less than zero (Study 1) or were unable to classify, i.e., d-prime was not significantly different from zero (Study 3, NoKnowledge group; Study 4, Preselected condition) the two stimulus categories. In other words, the GAN faces were either considered more real or as real as the REAL faces. If participants were told that there were fake faces present (Study 4, Knowledge group) or faces with flaws were included (Study 4, Flawed condition) in the experiment, then participants' performance in classifying stimulus types was significantly improved, i.e., d-prime was significantly larger than zero (Figure 10). The decision criterion remained stable across experiments (i.e., participants tended to say that faces were *real* in general), except in the Flawed condition of Study 4, where this bias was not significantly different from zero (Figure 11).

These results could stem from the rigorous pre-selection process we applied to the stimulus set rather than reflecting an inherent characteristic of GAN faces. To validate these findings with highly realistic examples of GAN technology, in Study 4 we investigated whether we would obtain the same results when participants are exposed to randomly sampled GAN and REAL faces from a pool of more and less realistic-looking stimuli. Our results indicated no such bias, and if participants were exposed to more obviously flawed GAN faces, they considered REAL faces to be more real than GAN faces, as one would expect from the literature investigating computer-generated faces (Balas and Horski, 2012; Carlson et al., 2012; Kätsyri, 2018; Matheson and McMullen, 2011). We specifically preselected the best examples that the algorithm can generate for two reasons. First, because only highly realistic faces are most likely to be used for marketing or by actors with malicious intent that aim to deceive users; these are thus most relevant to our current social climate. Second, because GAN technology continues to advance at such a rapid pace, a study of

flawed GAN faces would quickly become obsolete because updated algorithms would leapfrog and correct for such flaws, as already observed (Vincent, 2018).

In all studies, the age of the participants interacted with StimulusType, suggesting that older participants tended to say that GAN faces were more real than the REAL faces, and did so more often than younger participants (see Figures S4 and S5; See also Table S17). This pattern highlights the ways in which *digital natives* (people born after 1980) and *digital immigrants* (people born before 1980) may engage differently with technology, new media and the types of information conveyed (Prensky, 2001). Although it makes sense to consider digital natives and immigrants along a continuum rather than as a rigid dichotomy, digital fluency and engagement with information technology can be influenced by a range of factors besides age, also including other demographic characteristics, as well as psychological factors, social influence, and actual use of digital technologies (Wang et al., 2013). For the present studies we did not collect any data related to these other factors. Our findings are also broadly consistent with recent research showing that older adults are especially susceptible to misinformation and fake news, for reasons that have to do with their digital (il)literacy (Brashier and Schacter, 2020) and diminished ability to detect deception (Ruffman et al., 2012; Stanley and Blanchard-Fields, 2008). Given how gullible humans are, as research on desirability (Tappin et al., 2017) and confirmation biases showcases (Nickerson, 1998), advances in information technology and its capability for micro-targeting makes the further investigation of individual differences essential for understanding how humans can become more resilient to online misinformation and misleading content.

Limitations of the study

A relevant limitation of our study is that we could not fully characterize the variance observed in the participants' judgment and conformity. Our main aim was to investigate how GAN faces are perceived in relation to REAL faces and the behavioral consequences associated with it, but we also wanted to explain the reasons behind participants misclassifying the faces, either because of individual characteristics or image properties. We collected questionnaires and used indices (AQ, PI, TAS; see Section 2.7.1. of Data S1 file) to individuate three potential individual characteristics and explain some of the observed variance in the judgment, but they seemed not to have a significant influence on it (we only observed that TAS positively influenced the confidence). For the image properties, we observed that the attractiveness seems to have an important influence on realness perception, because this was significant across the studies for the realness task, but was not consistent for the conformity task. The expressiveness and the trustworthiness were not significant predictors either.

We had not predicted that the age and the gender of the participants could play an important role in the participants' judgment, but we did observe an interaction between the age and the stimulus type that we believe is worth investigating in future studies, because certain categories of participants might be more vulnerable to new technologies.

Conclusion

In general, we tend to over-attribute animacy to our surrounding world, including to inanimate objects, for evolutionary advantageous reasons (Guthrie, 1993). Animacy judgments on faces at least partly rely on the holistic processing of the whole face (Tanaka and Farah, 1993; Young et al., 1987), rather than individual features. For decades the literature on computer-generated faces struggled with the uncanny valley effect (Geller, 2008), but today's GAN faces are clearly past that point, as our evidence suggests. In addition to technological advances, the cultural context is also shifting, because virtual characters exert an ever expanding social and cultural influence through social media, as the example of Lil Miquela and her 2.5 million followers can attest. Moreover, from 2016 onwards, the introduction of the term "alternative facts", the spread of fake news in social media (Vosoughi et al., 2018), and the ever-increasing use of bots in political and social campaigns (Llewellyn et al., 2019; Shao et al., 2018) have shifted the socio-political landscape from being primarily truthful to being potentially deceptive. This current context of "fake news" seems to counteract our truth-default state that is our tendency to believe others (Levine, 2014). The potential for widespread activity of "fake agents" poses the question of how much their presence and activity can alter our truth-default state, eventually eroding social trust. As our Study 3 indicates, possessing knowledge about the presence of fake agents does decrease conformity and trust. This knowledge in itself seems to have a positive effect, making people more suspicious and reluctant to trust in an environment where fake agents operate. However, our research reveals that there are situations in which these fake agents are also the ones more likely to be perceived as real, hence trustworthy, pointing to the complex social consequences that generative technology and its (mis)use may have.

Against this cultural and technological background, several new questions arise for future investigations into the types of social information conveyed by GAN faces, such as intelligence, reputation and how they influence viewers across different social contexts. In times when the ground-truth of facts is questioned and our capacity for the production and manipulation of images is exponentially augmented, scientific efforts aimed at restraining the spread of misinformation should continue to focus on how we perceive realness and truthfulness. Given the accelerating development of such imaging technologies, their growing presence in the media and the current culture of misinformation in society, such questions should spur interdisciplinary investigations and provide data relevant for policy making.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants of study 1
  - Participants of study 2
  - Participants of study 3
  - Participants of study 4
- METHOD DETAILS
  - Stimuli generation and selection
  - Methods summary
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Analyses for the realness task (study 1, 3, and 4)
  - Analyses for the Conformity task (study 2 and 3)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105441.

ACKNOWLEDGMENTS

Manos Tsakiris was supported by the NOMIS Foundation Distinguished Scientist Award for the 'Body & Image in Arts & Science' project and a NOMIS Foundation Grant for the Centre for the Politics of Feelings

AUTHOR CONTRIBUTIONS

M.T., R.T., and N.V. provided the initial idea. R.T. and N.V. designed the experiments. R.T. and N.V. carried out computer programming. N.V. carried out the experiments. R.T. analyzed data. R.T. drafted the initial versions of the manuscript. R.T., N.V, S.C. and M.T. completed writing the paper. All contributed to the conceptual analysis of the results. All authors approved the final manuscript for submission.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure that the study questionnaires were prepared in an inclusive way. One or more of the authors of this article self-identifies as an underrepresented ethnic minority in science.

Received: March 16, 2022
Revised: July 12, 2022
Accepted: October 20, 2022
Published: December 7, 2022

REFERENCES

Abraham, A. (2015). How social dynamics shape our understanding of reality. In Neuroscience in Intercultural Contexts (International and Cultural Psychology), pp. 243–256. https://doi.org/10.1007/978-1-4939-2260-4_10.

Andrighetto, L., Baldissarri, C., Gabbiadini, A., Sacino, A., Rosa Valtorta, R., and Volpato, C. (2018). Social Influence Objectified Conformity: Working Self-Objectification Increases Conforming Behavior. https://doi.org/10.1080/15534510.2018.1439769.

Anobile, G., Cicchini, G.M., and Burr, D.C. (2014). Separate mechanisms for perception of numerosity and density. Psychol. Sci. *25*, 265–270. https://doi.org/10.1177/0956797613501520.

Avidan, G., Tanzer, M., and Behrmann, M. (2011). Impaired holistic processing in congenital prosopagnosia. Neuropsychologia *49*, 2541–2552. https://doi.org/10.1016/j.neuropsychologia.2011.05.002.

Bagby, R.M., Parker, J.D., and Taylor, G.J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. J. Psychosom. Res. *38*, 23–32. https://doi.org/10.1016/0022-3999(94)90005-1.

Balas, B., and Horski, J. (2012). You can take the eyes out of the doll, but. Perception *41*, 361–364. https://doi.org/10.1068/p7166.

Balas, B., and Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. Comput. Human Behav. *77*, 240–248. https://doi.org/10.1016/j.chb.2017.08.045.

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

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

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. J. Autism Dev. Disord. *31*, 5–17. https://doi.org/10.1023/A:1005653411471.

Behrmann, M., Thomas, C., and Humphreys, K. (2006). Seeing it differently: visual processing in autism. Trends Cogn. Sci. *10*, 258–264. https://doi.org/10.1016/j.tics.2006.05.001.

Ben-Shachar, M., Lüdecke, D., and Makowski, D. (2020). Effectsize: estimation of effect size indices and standardized parameters. J. Open Source Softw. *5*, 2815. https://doi.org/10.21105/joss.02815.

Beridze, I., and Butcher, J. (2019). When seeing is no longer believing. Nat. Mach. Intell. *1*, 332–334. https://doi.org/10.1038/s42256-019-0085-5.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., and White, J.S.S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. *24*, 127–135. https://doi.org/10.1016/j.tree.2008.10.008.

Brashier, N.M., and Schacter, D.L. (2020). Aging in an era of fake news. Curr. Dir. Psychol. Sci. *29*, 316–323. https://doi.org/10.1177/0963721420915872.

Brooks, M., Kristensen, K., Benthem, K., Magnusson, A., Berg, C., Nielsen, A., Skaug, H., Mächler, M., and Bolker, B. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. *9*, 378–400. https://doi.org/10.32614/rj-2017-066.

Carlson, C.A., Gronlund, S.D., Weatherford, D.R., and Carlson, M.A. (2012). Processing differences between feature-based facial composites and photos of real faces. Appl. Cogn. Psychol. *26*, 525–540. https://doi.org/10.1002/acp.2824.

Castelli, L., Zecchini, A., Deamicis, L., and Sherman, S.J. (2005). The impact of implicit prejudice about the the elderly on the reaction to stereotype confirmation and disconfirmation. Curr. Psychol. *24*, 134–146.

Castelli, L., Vanzetto, K., Sherman, S.J., and Arcuri, L. (2001). The explicit and implicit perception of in-group members who use stereotypes: blatant rejection but subtle conformity. J. Exp. Soc. Psychol. *37*, 419–426. https://doi.org/10.1006/jesp.2000.1471.

Castelli, L., Arcuri, L., and Zogmaister, C. (2003). Perceiving ingroup members who use stereotypes: implicit conformity and similarity. Eur. J. Soc. Psychol. *33*, 163–175. https://doi.org/10.1002/ejsp.138.

Demoulin, S., Torres, R.R., Perez, A.R., Vaes, J., Paladino, M.P., Gaunt, R., Pozo, B.C., and Leyens, J.-P. (2004). Emotional prejudice can lead to infra-humanisation. Eur. Rev. Soc. Psychol. *15*, 259–296. https://doi.org/10.1080/10463280440000044.

Geller, T. (2008). Overcoming the uncanny valley. IEEE Comput. Graph. Appl. *28*, 11–17. https://doi.org/10.1109/MCG.2008.79.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. Commun. ACM *63*, 139–144. https://doi.org/10.1145/3422622.

Gray, H.M., Gray, K., and Wegner, D.M. (2007). Dimensions of mind perception. Science *315*, 619. https://doi.org/10.1126/science.1134475.

Green, R.D., MacDorman, K.F., Ho, C.-C., and Vasudevan, S. (2008). Sensitivity to the proportions of faces that vary in human likeness. Comput. Human Behav. *24*, 2456–2474. https://doi.org/10.1016/j.chb.2008.02.019.

Guthrie, S. (1993). Faces in the Clouds: A New Theory of Religion (Oxford University Press).

Haxby, J.V., Hoffman, E.A., and Gobbini, M.I. (2000). The distributed human neural system for face perception. Trends Cogn. Sci. *4*, 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0.

Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., and Traue, H.C. (2010). Expression intensity, gender and facial emotion recognition: women recognize only subtle facial emotions better than men. Acta Psychol. *135*, 278–283. https://doi.org/10.1016/j.actpsy.2010.07.012.

Jessimer, M., and Markham, R. (1997). Alexithymia: a right hemisphere dysfunction specific to recognition of certain facial expressions? Brain Cogn. *34*, 246–258. https://doi.org/10.1006/brcg.1997.0900.

Judd, C.M., Westfall, J., and Kenny, D.A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. J. Pers. Soc. Psychol. *103*, 54–69. https://doi.org/10.1037/a0028347.

Karras, T., Laine, S., and Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks.

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. Front. Psychol. *9*, 1362–1412. https://doi.org/10.3389/fpsyg.2018.01362.

Kress, T., and Daum, I. (2003). Developmental prosopagnosia: a review. Behav. Neurol. *14*, 109–121. https://doi.org/10.1155/2003/520476.

Levine, T.R. (2014). Truth-default theory (TDT): a theory of human deception and deception detection. J. Lang. Soc. Psychol. *33*, 378–392. https://doi.org/10.1177/0261927X14535916.

Llewellyn, C., Cram, L., Hill, R.L., and Favero, A. (2019). For whom the bell trolls: shifting troll behaviour in the twitter brexit debate. J. Commun. Media Stud. *57*, 1148–1164. https://doi.org/10.1111/jcms.12882.

Loeys, T., Moerkerke, B., De Smet, O., and Buysse, A. (2012). The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. Br. J. Math. Stat. Psychol. *65*, 163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x.

Looser, C.E., and Wheatley, T. (2010). The tipping point of animacy. Psychol. Sci. *21*, 1854–1862. https://doi.org/10.1177/0956797610388044.

Lüdecke, D., Ben-Shachar, M., Patil, I., and Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. J. Open Source Softw. *5*, 2445. https://doi.org/10.21105/joss.02445.

Lüdecke, D. (2018). Ggeffects: tidy data frames of marginal effects from regression models. J. Open Source Softw. *3*, 772. https://doi.org/10.21105/joss.00772.

Macmillan, N.A., and Creelman, C.D. (2004). Detection Theory: A User's Guide.

Makowski, D., Ben-Shachar, M.S., Patil, I., and Lüdecke, D. (2021). Automated Results Reporting as a Practical Tool to Improve Reproducibility and Methodological Best Practices Adoption (CRAN).

Matheson, H.E., and McMullen, P.A. (2011). A computer-generated face database with ratings on realism, masculinity, race, and stereotypy.

Behav. Res. Methods *43*, 224–228. https://doi.org/10.3758/s13428-010-0029-9.

Mcdonald, K. (2018). How to recognize fake AI-generated images. Medium. https://kcimc.medium.com/how-to-recognize-fake-ai-generated-images-4d1f6f9a2842.

Nickerson, R.S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. Rev. Gen. Psychol. *2*, 175–220.

Nightingale, S.J., and Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc. Natl. Acad. Sci. USA *119*. e2120481124. https://doi.org/10.1073/pnas.2120481119.

Paladino, M.P., Mazzurega, M., Pavani, F., and Schubert, T.W. (2010). Synchronous multisensory stimulation blurs self-other boundaries. Psychol. Sci. *21*, 1202–1207. https://doi.org/10.1177/0956797610379234.

Palermo, R., and Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. Neuropsychologia *45*, 75–92. https://doi.org/10.1016/j.neuropsychologia.2006.04.025.

Parello-Plesner, J. (2018). China's LinkedIn honey traps - the american interest. Am. Interes. https://www.the-american-interest.com/2018/10/23/chinas-linkedin-honey-traps/

Pendry, L.F., and Macrae, C. (1994). Stereotypes and mental life - the case of the motivated but thwarted tactician. J. Exp. Soc. Psychol. *30*, 303–325.

Prensky, M. (2001). Digital natives, digital immigrants Part 1. Horizon *9*, 1–6. https://doi.org/10.1108/10748120110424816.

R Core Team (2022). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing). https://www.R-project.org/.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. Annu. Rev. Psychol. *57*, 199–226. https://doi.org/10.1146/annurev.psych.57.102904.190208.

Rothman, J. (2018). In the age of A.I., is seeing still believing? (New Yorker).

Ruffman, T., Murray, J., Halberstadt, J., and Vater, T. (2012). Age-related differences in deception. Psychol. Aging *27*, 543–549. https://doi.org/10.1037/a0023380.

Satter, R. (2019). Experts: Spy used AI-generated face to connect with targets. APnews.com. https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

Seyama, J., and Nagayama, R.S. (2009). Probing the uncanny valley with the eye size aftereffect. Presence. (Camb). *18*, 321–339. https://doi.org/10.1162/pres.18.5.321.

Shah, P., Gaule, A., Sowden, S., Bird, G., and Cook, R. (2015). The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. R. Soc. Open Sci. *2*, 140343. https://doi.org/10.1098/rsos.140343.

Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. Nat. Commun. *9*, 4787. https://doi.org/10.1038/s41467-018-06930-7.

Sofer, C., Dotsch, R., Wigboldus, D.H.J., and Todorov, A. (2015). What is typical is good: the influence of face typicality on perceived trustworthiness. Psychol. Sci. *26*, 39–47. https://doi.org/10.1177/0956797614554955.

Stanley, J.T., and Blanchard-Fields, F. (2008). Challenges older adults face in detecting deceit: the role of emotion recognition. Psychol. Aging *23*, 24–32. https://doi.org/10.1037/0882-7974.23.1.24.

Stephen, I.D., Law Smith, M.J., Stirrat, M.R., and Perrett, D.I. (2009). Facial skin coloration affects perceived health of human faces. Int. J. Primatol. *30*, 845–857. https://doi.org/10.1007/s10764-009-9380-z.

Tanaka, J.W., and Farah, M.J. (1993). Parts and wholes in face recognition. Q. J. Exp. Psychol. *46*, 225–245. https://doi.org/10.1080/14640749308401045.

Tappin, B.M., van der Leer, L., and McKay, R.T. (2017). The heart trumps the head: DESIRABILITY bias in political belief revision. J. Exp. Psychol. Gen. *146*, 1143–1149. https://doi.org/10.1037/xge0000298.

Todorov, A., Olivola, C.Y., Dotsch, R., and Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. Annu. Rev. Psychol. *66*, 519–545. https://doi.org/10.1146/annurev-psych-113011-143831.

Toelch, U., and Dolan, R.J. (2015). Informational and normative influences in conformity from a neurocomputational perspective. Trends Cogn. Sci. *19*, 579–589. https://doi.org/10.1016/j.tics.2015.07.007.

Vaes, J., Paladino, M.P., Castelli, L., Leyens, J.-P., and Giovanazzi, A. (2003). On the behavioral consequences of infrahumanization: the implicit role of uniquely human emotions in intergroup relations. J. Pers. Soc. Psychol. *85*, 1016–1034. https://doi.org/10.1037/0022-3514.85.6.1016.

Vaes, J., Paladino, M.P., and Magagnotti, C. (2011). The human message in Politics: the impact of emotional slogans on subtle conformity. J. Soc. Psychol. *151*, 162–179. https://doi.org/10.1080/00224540903510829.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. Q. J. Exp. Psychol. *43*, 161–204. https://doi.org/10.1080/14640749108400966.

van Cappellen, P., Corneille, O., Cols, S., and Saroglou, V. (2011). Beyond mere compliance to authoritative figures: religious priming increases conformity to informational influence among submissive people. Int. J. Psychol. Relig. *21*, 97–105. https://doi.org/10.1080/10508619.2011.556995.

Vincent, J. (2018). These faces show how far AI image generation has advanced in just four years. Verge.

Vincent, J. (2020). An Online Propaganda Campaign Used AI-Generated Headshots to Create Fake Journalists (Verge.com). https://www.theverge.com/2020/7/7/21315861/ai-generated-headshots-profile-pictures-fake-journalists-daily-beast-investigation.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. Science *359*, 1146–1151. https://doi.org/10.1126/science.aap9559.

Wang, Q., Com Michael Myers, B.D., Sundaram, D., and Buxmann, B.Y. (2013). BISE-STATE of the ART Digital Natives and Digital Immigrants towards a Model of Digital Fluency the Authors. https://doi.org/10.1007/s12599-013-0296-y.

Webster, R., Rabin, J., Simon, L., and Jurie, F. (2021). This person (probably) exists. Identity Membership Attacks Against GAN Generated Facespreprint at arXiv. https://doi.org/10.48550/arXiv.2107.06018.

Weigelt, S., Koldewyn, K., and Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: a review of behavioral studies. Neurosci. Biobehav. Rev. *36*, 1060–1084. https://doi.org/10.1016/j.neubiorev.2011.12.008.

Young, A.W., Hellawell, D., and Hay, D.C. (1987). Configurational information in face perception. Perception *16*, 747–759. https://doi.org/10.1068/p160747.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Preprocessed data	This paper	https://osf.io/9t3sy/files/osfstorage/Scripts/DS_AllStudies.R
Stimuli	This paper	https://osf.io/9t3sy/files/osfstorage/Stimuli.rar
Example of stimuli anomaly	This paper	https://osf.io/9t3sy/files/osfstorage/Examples of stimuli anomalies.pdf
Software and algorithms		
R version 4.1.1 (2021-08-10)	R Project	https://www.r-project.org/
Prolific	Prolific	Prolific.co
Qualtrics	Qualtrics	Qualtrics.com
Gorilla	Gorilla	Gorilla.sc
Matlab 2019b	Mathworks, Inc	Mathworks.com
Gimp 2.10.18	Gimp	Gimp.org
R Studio version 2021.09.0 Build 351	RStudio	https://www.rstudio.com/
Scripts for analyses	This paper	https://osf.io/9t3sy/files/osfstorage/Scripts/A03_MainAnalyses_iScience.Rmd
Details of the analysis	This paper	https://osf.io/9t3sy/files/osfstorage/GAN_AnalysesList_iScience_OSF.xlsx
Software used	This paper	https://osf.io/9t3sy/files/osfstorage/Software.pdf
Other		
REAL faces	Flickr-Faces-HQ	github.com/NVlabs/ffhq-dataset
GAN faces	This Person Does Not Exist	thispersondoesnotexist.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Raffaele Tucciarelli (rtucciarelli@gmail.com).

Materials availability

This study did not generate new materials.

Data and code availability

- All data and statistical analyses that support the findings of this study are publicly available in Open Science Framework at https://osf.io/9t3sy/files/osfstorage.
- Preprocessed data have been deposited at OSF.io and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- Stimuli have been deposited at OSF.io and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- Examples of stimuli anomaly have been deposited at OSF.io and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at OSF.io and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Details of the analysis have been deposited at OSF.io and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.

- A list of the software used has been deposited at OSF.io and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact on request.

Correspondence with questions and requests for materials should be addressed to R.T. (rtucciarelli@gmail.com) or M.T. (Manos.Tsakiris@rhul.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The four main studies were conducted online by recruiting human participants using the Prolific platform (Prolific.co). All studies were approved by the Royal Holloway, University of London, Ethics Committee. Participants were informed about the study and gave their written consent before the beginning of the experiment.

Participants of study 1

An *a priori* power analysis using pilot data (section 2.5 of the Data S1 file) determined the recruitment of at least 116 participants. We began by collecting 122 participants to account for exclusions. Six were excluded beacuse of failed attention checks or too many missing trials; nine people were excluded because they responded *real* to all pictures. Although these can be genuine responses, these participants were removed because they would not add any informative variance to the analysis. The final sample consisted of 107 individuals (58 women, 48 men; one participant chose the "Other" option; age: M = 28.18, SD = 9.62).

Participants of study 2

We invited the same participants who already took part in Study 1, to participate in our conformity study (Study 2). Seventy-three of them responded to our invitation. Nine participants were excluded because of failed attention checks or too many missing trials, thus our sample consisted of 64 participants. Subsequently, three participants were further removed because they always (i.e., for all trials) copied the number at the top of the screen (which is the estimated number of letters displayed, supposedly provided by the person they saw on the photo) instead of attempting to provide their own estimate. In addition, 10 participants were removed because they were previously excluded from Study 1 as outliers. Following this, the final analyses were conducted on 55 participants (31 women; age: M = 29.1, SD = 9.86) in total. Participants (N = 8) were excluded if they failed two or more attention checks (e.g., instruction to respond by typing a specific number) or missed more than 5% of the trials.

Participants of study 3

Participants were divided into two groups: *Knowledge* and *NoKnowledge*. The subsequent analyses were conducted on 232 participants (81 women; one participant chose the option "other"; age: M = 26.38, SD = 9.80), 116 per group. Participants first performed the Conformity task (as in Study 2) followed by the Realness task (as in Study 1). They were given the definitions of Real and Fake faces (as in Study 2) at different stages of the experiment depending on the group. Participants were not informed of the proportions of each category of faces but were incentivized with a bonus for accuracy.

Participants of study 4

We recruited a sample of 122 participants. After rejections because of failed attention checks, we retained 116 participants (38 women; age: M = 24.84, SD = 7.59). Participants were excluded if they failed two or more attention checks or missed more than 5% of the trials. The study was hosted on Gorilla platform (gorilla.sc).

Participants' summary

Study	Group	N	Mean Age (SD)	Tasks	Main analysis
Study 1		107 (58 women)	28.18 (9.62)	1. Realness task 2. Questionnaires	GLMM Logistic (Binomial distribution)
Study 2	[Subgroup of Study 1]	55(31 women)	29.1 (9.86)	Conformity task	GLMM hurdle (Negative binomial distribution)
Study 3	Knowledge	116 (44 women)	26.71 (10.2)	1. Conformity Task 2. Realness task	1. GLMM hurdle 2. GLMM Logistic
	NoKnowledge	116 (37 women)	26.09 (9.04)	1. Conformity Task 2. Realness task	1. GLMM hurdle 2. GLMM Logistic
Study 4		116 (38 women)	24.84 (7.59)	Realness Task	GLMM Logistic

METHOD DETAILS

Stimuli generation and selection

The algorithms which produce generative adversarial network (GAN) faces are being developed and upgraded at a very fast pace. Supplying increasingly more realistic-looking faces, they could be used to fool people into assuming that GAN faces depict real people. With this in mind, we aimed to investigate the most advanced examples of faces that the algorithm can currently produce, to avoid our findings becoming obsolete with its development.

When selecting GAN faces to use in our studies, we included faces without obvious rendering errors (e.g., see Mcdonald, 2018); for example, blurry teeth, unnatural asymmetries, not well rendered objects (e.g., glasses, earrings, etc.), which are clearly visible to the naked eye. This was done to minimize the role of salient flaws, because people's ability to detect obvious computer-generated images as fake has already been established elsewhere (Balas and Horski, 2012; Carlson et al., 2012; Kätsyri, 2018; Matheson and McMullen, 2011). This study thus investigated how we process state-of-the-art GAN faces, produced by StyleGAN algorithm (thispersondoesnotexist.com) in comparison to REAL faces which were taken from the Flickr-Faces-HQ (available at github.com/NVlabs/ffhq-dataset). We chose these real faces because they were the source images used to train the algorithm, possessing the same biases, and are thus most closely comparable. For further details on the stimulus generation and selection, please refer to sections 2.1, 2.2 and 8 of the Data S1 file.

Before Study 1 we conducted three Stimuli Validation studies, each estimating the intensity of a specific trait expressed in each of the faces we used. We asked three different groups of participants to estimate how attractive ($N_A = 31$), trustworthy ($N_T = 30$) and expressive ($N_E = 30$) they find both REAL and GAN faces, respectively. For more details, see section 2.4 of the Data S1 file and Figure S1 of the Supplemental information.

Methods summary

The two main dependent variables of interest across studies were: 1) the categorical Judgments (i.e., *real/fake*) to the faces in the Realness tasks; and 2) the *Conformity Index* scores, indicating the amount of conformity to a face, collected using a Conformity task. In Study 1, we asked participants to take part in a Realness task. After this task, they were also asked to complete three questionnaires (see next paragraph questionnaires (Study 1)). In Study 2, a subset of participants that took part in Study 1 were asked to also participate in the Conformity task. Finally, in Study 3, two new groups of participants (for Knowledge and NoKnowledge condition respectively) were first asked to do the Conformity task and then the Realness task. All studies were designed and hosted on Qualtrics (qualtrics.com) or Gorilla (gorilla.sc). Participants were recruited via Prolific (prolific.co). The three studies were pre-registered (Study 1: https://osf.io/5hswy; Study 2: https://osf.io/hae8q; Study 3: https://osf.io/x85pr). Subsequently, we ran Study 4 in which participants took part adapted version of the Realness task that aimed at validating our stimuli selection and our results.

Realness Task (Study 1, 3, 4)

Participants were instructed to judge if a face was real or fake. The instructions defined what was meant by a real ("images depicting genuine, unaltered faces of people") or a fake face ("the face has been

artificially generated by an algorithm and does not depict an existing person"). One hundred faces (50 GAN and 50 REAL) were randomly presented on the screen with a white background. Each picture was presented for 3 s only. This exposure time was chosen as a compromise to allow participant to see the face for a sufficient amount of time, but not for too long, because we wanted to trigger a response for a first impression view of the face. Participants then answered two-alternative forced choice (2AFC) questions with a key press and indicated their confidence by moving a slider on a continuous scale (from 0 "Low confidence/Guess" to 100 "High confidence/Certain"). Four catch trials were used as attention checks. Participants who failed two or more attention checks were excluded from the analysis. For Study 4, the procedure for the participants was identical except that the stimuli also included 100 faces (50 GAN and 50 REAL) that were discarded during the stimulus selection procedure (see section 2.1 and 2.10 of the Data S1 file).

Questionnaires (Study 1)

After the main task and in Study 1 only, participants completed the Autism Spectrum Quotient (Baron-Cohen et al., 2001), the Prosopagnosia index (Shah et al., 2015), and the Toronto Alexithymia Scale (Bagby et al., 1994). We selected these traits given their role in face processing (Avidan et al., 2011; Behrmann et al., 2006; Kress and Daum, 2003; Weigelt et al., 2012) and social cognition (Haxby et al., 2000; Jessimer and Markham, 1997) (for details see section 2.7 of the Data S1 file). They also answered two questions regarding their prior experience and knowledge of the existence of GAN faces (*yes/no*) and in what capacity they had these experiences, if they answered "yes". We added this to account for possible effect of different exposure levels on realness judgments. For details and reasons for choosing these questionnaires, see section 2.7 of the Data S1 file.

Conformity task (Study 2 and 3)

Participants took part in a modified version of the number of letter estimation task (Castelli et al., 2001). The procedure of Study 2 is schematically summarized in Figure 4. Participants had to estimate the number of letters presented on the screen at different densities. Importantly, at each trial and before each density display, they saw one of the faces used in Study 1 (either a GAN or a REAL face) for 3 s; furthermore, after the face disappeared and on top of the letter display, they were also provided with the estimate given by that face for the same letter display. Participants decided if they wanted to use such an estimate as an anchor to provide their own estimate. They were informed that "*real faces tried very hard to provide a correct estimate*", and therefore the number could be informative, but the number provided by GAN faces was randomly generated by an algorithm, and therefore was not necessarily informative. Therefore, we implicitly asked participants to judge whether they believed a face was *real* or *fake*. To incentivize participants to use all available information, we rewarded their accuracy on a random trial with a bonus. The letter displays were generated using Matlab 2019b. The monitor was a Dell UP2516D (55.29 × 31.10 cm; 2560 × 1440 resolution; 0.216 × 0.216 mm pixel pitch). Each display consisted of 200 letters *As* randomly distributed within an imaginary circle. We varied the radius to generate 500 displays with 5 different densities to manipulate the numerosity perception (Anobile et al., 2014). More specifically, 5 radii varied from 39 to 78 pixels (with equal step). For each radius we generated 100 random displays by sampling the letter positions as Cartesian coordinate pairs from a normal distribution with mean zero and standard deviation equal to the radius. To account for extreme locations, we repeated the sampling until all letter positions were within 3 standard deviations from the mean of all position radii. The density of each display was computed as the number of letter within a unit area divided by the unit area. Therefore, for each radius a letter could have appeared at any angle of the circle and at any distance between the centre of the circle and the radius. In total, we generated 500 displays (100 displays per radius) from where a display was randomly sampled at each trial (N = 100 trials). The estimate on top of the letter display, that was supposedly given by the face displayed just before, was actually a random number generated from a normal distribution with a mean of 200 (i.e., the number of actual dots in the display) and a standard deviation of 25. These values were chosen to match the value used in Castelli et al. (2001). The sampled numbers were then rounded to the nearest integer because the number of dots is an integer. The task involved seeing 100 faces (50 GAN and 50 REAL) and 4 attention check trials which were already included in Study 1, each for 3 s and had 10 s to provide their estimate of the number of letters on the screen. It was divided into two blocks, with a 3-min break in between. In Study 3, participants were divided into the *Knowledge* and the *NoKnowledge* groups. Both groups started with the Conformity task followed by the Realness task, but differed on the type of information that they received at the beginning of the Conformity task. The Knowledge group was first given the same definitions as in Study 1 followed by a paragraph to provide a social context: "Such technology of

generating artificial faces of non-existing people can be used in various useful contexts (e.g. improve the quality of old photos, generate new model images for commercial websites, etc.), but is also being used with malicious intentions, such as generate fake social media profiles that could influence social and political behavior. It is therefore important and timely to investigate how we process such faces". This was added to reinforce the importance of the distinction in Realness and to effectively frame the Conformity task. Participants were told that some of the faces had put a lot of effort to provide a good estimate of the number of letters on the screen, others just responded randomly and that they had to "Use any information available to you to consider whether a face has given an informative response that you may or may not want to take on board". The sentence intended to heighten the participants' awareness and effort in gathering information from non-verbal cues of the faces, but was purposefully vague about what "cues" meant. This intended to increase ecological validity, where people might be aware of the existence of fake faces (e.g. on social media), but are not told that the faces they are looking at might not be real. The participants then proceeded to the Realness task. The NoKnowledge group did not receive any indication about the nature of the presented faces in the initial Conformity task. The only piece of information provided was that some of the faces on the images tried to make good estimates, whereas others just responded randomly and that they should use all the information available to them, when estimating the number of letters on the screen. After the Conformity task, we again introduced the definitions of *Real* and *Fake* faces with social context and proceeded with the Realness task (see section 2.9 of the Data S1 file for further details).

QUANTIFICATION AND STATISTICAL ANALYSIS

You can see the planned analyses (see section *Analysis Plan*) for Study 1 at https://osf.io/5hswy; for Study 2 at https://osf.io/hae8q; and for Study 3 at https://osf.io/x85pr. We strictly followed the analysis plan that what we stated in our pre-registrations, but please note that we realized that some of the analyses that we originally planned were actually less appropriate for our types of data. We therefore added equivalent but more appropriate analyses that mainly involved the use of generalized linear mixed models (GLMM). We explicitly say when we use planned analyses or exploratory analyses through the manuscript. Also, for transparency and to fulfill the pre-registration, we report and discuss the outcome of the planned but inappropriate analyses in the Data S1 file (section 2.6).

Analyses for the realness task (study 1, 3, and 4)

The main aim of this analysis was to identify the best predictors of the participant's judgment, meaning saying whether a face was *real* or *fake*. Our main focus was on the StimulusType (either a REAL or a GAN faces), therefore one of our questions was whether we could predict the response of participants based on the presented face while controlling for other potentially confounding factors. Because the experimental design required the presentation of different exemplars (i.e., individual faces) of a same category (i.e., either GAN or REAL face) to each participant, we used *generalized linear mixed models* (GLMM) that included the participants and the images as random effects to control for variations related to these two components. This is a good procedure when one wants to draw inferences that generalize across participants and stimuli avoiding the biases because of averaging, inflating Type I Error (Judd et al., 2012). Because we also planned ANOVAs to analyze these data, our main dependent variable was the participants' judgment (either *real* or *fake*) which is a dichotomous variable. We therefore used a GLMM Logistic regression analysis with a *binomial* distribution. Our main model included the *StimulusType* (REAL|GAN), the *Stimulus Gender* (Male|Female), the participants' *Age* and *Gender*, and the interaction between the *StimulusType* and the participants' *Age* and *Gender*; but we also ran models that included characteristics of the faces (Attractiveness, Expressiveness, and Trustworthiness) and, for Study 1 only, the indices from the three questionnaires (AQ, PI, TAS). For Study 3, we additionally included the *Group* (Knowledge|NoKnowledge) as a predictor for the participants' judgment and its interaction with the *StimulusType* and the *Stimulus Gender*. To further characterize the latter interactions, we fitted two models using the two group dataset separately. As *random-effect* components, we included variation in the intercept because of participants (Participant ID) and images (Image ID). The R version of these models can be seen in the Excel file GAN_AnalysesList_iScience_OSF.xlsx at https://osf.io/9t3sy/files/osfstorage. Estimated parameters will be provided as Odd Ratio (OR), which quantify the odds that a parameter level can predict the outcome relative to a baseline level. An OR = 1 indicates that the parameter is not significant.

Exploratory analysis

We computed two main indices to quantify the realness associated to an image (proportion of real responses associated to an image, or pRI) or to a stimulus category (proportion of real responses associated to a stimulus category, or pRC). For each image, pRI was computed by counting the number of participants that classified the image as *real* divided by the number of participants. For each participant, pRC was computed by counting the number of times a participant said that a stimulus category was real divided by the number of images in a category (N = 25 repetitions for each combination of StimulusType and Gender). Four one-sample t-tests were used to test the null hypothesis that the pRI means were 0.5, indicating that on average people responded randomly. Similarly, four one-sample t-tests were used to test the null hypothesis that the pRC means were 0.5, indicating that on average people responded randomly.

Analyses for the Conformity task (study 2 and 3)

For each participant, the amount of conformity associated to each face was quantified using the *Conformity Index*, that was computed as the absolute difference between the estimate number provided (i.e., the one on top of the letter array that indicated the estimate supposedly given by the current face) and the participants' responses (Castelli et al., 2001; Vaes et al., 2003). The main aim of these analyses was to identify the best predictors that could explain the observed variance of the *Conformity Index*. More specifically, in our pre-registration of Study 2, we originally hypothesized that the GAN faces would have been associated with smaller *Conformity Index* values, indicating higher levels of conformity, than REAL faces, because in Study 1 GAN faces were generally considered more *real* than REAL faces. In the pre-registration of Study 3, and based on the results of Study 2, we planned to also investigate the role of the participants' judgment in explaining the observed conformity levels. In fact, we reasoned that what it counts is just how participants would perceive the realness associated with an image, rather than the actual type of stimulus. Note that also for these analyses, we originally planned ANOVAs, but subsequently realized that these were less appropriate analyses to use, as discussed above (Judd et al., 2012). We therefore decided to use the GLMM regression analyses for the Conformity task, but we also report the results of the planned ANOVA in the Data S1 file (see sections 3.2.1. and 3.3.1.) for transparency.

The *Conformity Index* has a lower bound of zero, indicating maximal conformity. Therefore, small *Conformity Index* values suggest higher conformity, meaning that the observed face was trusted sufficiently to propose an estimate that was similar or identical to the one suggested by the face. We used a GLMM with a negative binomial distribution, or hurdle analysis (Loeys et al., 2012), to explore which predictors (i.e., *Judgment*, *StimulusType*, *Stimulus Gender*) could explain the observed Conformity Indices. The stimulus characteristics (*Attractiveness*, *Expressiveness*, *Trustworthiness*, *Stimulus Age*, *Face angle*, *Smile*), the participant *Age* and Gender, and the interactions between *Age* and *StimulusType*, *Stimulus Gender*, and *Judgment* were also used as confounding predictors. For Study 3 we also added the *Group* (Knowledge|NoKnowledge) as a predictor of interest. A negative binomial regression was used because the CI was positively skewed and bounded at zero (N = 415 zeros out of 5427). This indicated that violations of the multiple linear regression assumptions were likely (Bolker et al., 2009) and over-dispersion might have occurred, meaning that the variance of the distribution was likely to become larger than the mean as this increased (Loeys et al., 2012). Note that in our study, the zeros were not because of sampling (i.e., missing responses), but were informative of the underpinning process, indicating absolute conformity or just plain copying of the estimate. The hurdle analysis takes into account the presence of such zeros by defining two model components: 1) The *conditional* model that fits only values larger than zeros to explain conformity from high (small *Conformity Index* values) to low (high *Conformity Index* values); 2) The *zero-inflated* model that treats the data binary, indicating whether a participant would conform (*Conformity Index* = 0) or not (*Conformity Index*> 0). See also the results section of study 2 for details.

The Analyses were conducted in R 4.1.1 (2021-08-10; R Core Team, 2021) and Matlab 2019b (MathWorks, Inc.). The generalized mixed-models (GLMM) analyses were conducted using the glmmTMB package (Bolker et al., 2009; Brooks et al., 2017). Nested models were compared using the likelihood ratio test using the *anova* function. For summary and report of the fitting analysis, we used *parameter* (Lüdecke et al., 2020) and *report* (Makowski et al., 2021) packages. For plotting the effect plots, we used the *ggeffects* (Lüdecke, 2018) package. Effect sizes were computed using the *effectsize* (Ben-Shachar et al., 2020) package. For a complete list of the software used, please see the file *Software.pdf* in https://osf.io/9t3sy/files/osfstorage.