AI-generated faces are becoming more trustworthy

Alexis A. McGuire Department of Psychology, Lancaster University, Lancaster, UK

Maty Bohacek Department of Electrical Engineering, Stanford University, Stanford, CA, USA

Hany Farid School of Information, University of California, Berkeley, Berkeley, CA, USA

Paul J. Taylor Department of Psychology, Lancaster University, Lancaster, UK

Sophie J. Nightingale Department of Psychology, Lancaster University, Lancaster, UK

The development of generative artificial intelligence (GAI) has democratized access to high-quality image generation, creating both opportunities and risks. Research shows that generative adversarial networks (GANs) can generate faces that are nearly indistinguishable from and more trustworthy than real ones. However, the literature currently contains no direct comparison of the realism and trustworthiness of GAN-synthesized faces and faces produced by more recent models. Here we test whether faces generated by diffusion models (DMs), a new and more powerful GAI architecture, elicit trust and have similarly passed through the uncanny valley. In Experiment 1, participants ($N = 169$) saw 96 facial images and for each face indicated whether it was real or AI generated. In Experiment 2, participants ($N = 87$) again saw 96 facial images and this time were asked to rate how trustworthy each face appeared. We find that DMs produce less photo-realistic faces than GANs, but those faces are rated as more trustworthy than GANs and real faces. As the realism and availability of GAI continue to increase, it is now—more than ever—critical to understand this threat and develop strategies to mitigate potential harms to individuals, organizations, and democracies.

Introduction

Humans have evolved to become experts at processing real faces (Atkinson & Adolphs, 2011). Our perception of faces occurs automatically (Todorov, Said, Engell, & Oosterhof, 2008), and inferences can be made in as little as 100 milliseconds (Willis & Todorov, 2006). From an evolutionary perspective, we have acquired this skill to understand the identities of others and facilitate interactions, from communicating with friends to detecting threats for survival purposes (Öhman, Lundqvist, & Esteves, 2001). However, the rise of technological advancements presents challenges for human perception of faces—namely, the inception of sophisticated computational techniques that have made it possible to create highly realistic fake faces using artificial intelligence (AI). These AI-generated faces pose questions as to whether our expertise in the processing of real faces will translate to fake ones, presenting a need to understand the mechanisms behind human perception of AI-synthesized faces and whether this differs by synthesis model.

One of the earliest AI technologies used to produce faces with a close likeness to human faces is the generative adversarial network (GAN) (Goodfellow et al., 2014). From previous literature, we know that humans have a limited ability to detect GAN-synthesized faces, with multiple studies reporting near-chance performance, suggesting these faces are highly realistic (Bray, Johnson, & Kleinberg, 2023; Nightingale & Farid, 2022). Miller and colleagues provide further evidence that GAN-generated faces exhibit high photorealism and identify what they refer to as a "hyperrealism effect" (Miller et al., 2023). Using the White/Caucasian face stimuli from Nightingale and Farid (2022), participants rated whether 100 faces were "real" or "AI generated." The findings revealed

Citation: McGuire, A. A., Bohacek, M., Farid, H., Taylor, P. J., & Nightingale, S. J. (2026). AI-generated faces are becoming more trustworthy. *Journal of Vision*, *26*(7):3, 1–9, https://doi.org/10.1167/jov.26.7.3.

https://doi.org/10.1167/jov.26.7.3 Received November 13, 2025; published July 7, 2026 ISSN 1534-7362

that participants were significantly more likely to respond "real" for AI-generated faces than for real faces. Adding further concern, research has shown that people often demonstrate high confidence in their incorrect classifications of GAN-synthesized faces as real or AI—people's lack of insight into their own performance indicates there are limited cues within GAN-generated faces that the human perceptual system can pick up on (Bray et al., 2023; Miller et al., 2023). In addition, another study also found that the authenticity of GAN-generated faces was questioned less than real faces (Lago et al., 2021). It remains unknown, however, whether the findings of these two studies extrapolate beyond the White/Caucasian face stimuli to other racial groups. Overall, there is a growing body of evidence indicating that faces generated using a GAN architecture display high levels of perceptual realism. Research suggests that GAN-generated faces have now "passed through the uncanny valley," that is, the faces are so realistic that they can trick our perceptual system into processing them as real faces (Mori, MacDorman, & Kageki, 2012).

What is it, then, that makes these faces so realistic? Many psychological theories have attempted to underpin the mechanisms behind facial processing, and researchers have been motivated to understand what it is that makes these AI-generated faces appear human-like. One suggestion is that the high levels of realism could be linked to face space theory (Valentine, 1991). Face space is a well-accepted theoretical construct that provides a multidimensional space where mental representations of faces are encoded, based on their features—an individual's experience will determine the structure of their own mental representation of faces. The axes and dimensions of face space are created through learning based on the kinds of faces an individual sees most often. Facial features that are encountered frequently cluster together on each dimension and form a strong central point, which gives a representation of an "average face." When an individual sees a new face, it is assessed relative to this distribution. If the new face falls within or close to the clustered region, then it feels more typical or familiar. Hence, more average faces appear to be more familiar. In an attempt to understand the "hyperrealism effect," Miller et al. (2023) investigated which facial attributes participants associated with real and GAN-synthesized faces. The results showed that GAN-synthesized faces were rated as more average and more familiar than real faces, in line with predictions from face space theory. Again, a key limitation, however, is that the stimulus set only included White/Caucasian faces, and therefore, it is unclear whether these findings would hold for other racial groups. It seems, though, that the GAN architecture might generate faces that coincide with human expectations of what a "real face looks like."

Not only are GAN-synthesized faces highly realistic, but they are also rated as more trustworthy than real faces (Nightingale & Farid, 2022). Although somewhat speculative, researchers have hypothesized that the GAN synthesis process is likely biased toward generating faces close to the average of the faces it had access to in the training dataset (a hypothesis that gained some support from later research; Miller et al., 2023). Drawing on the wider psychological literature, more average faces tend to be rated as more trustworthy than distinct faces (Sofer, Dotsch, Wigboldus, & Todorov, 2015). Therefore, this averageness in the appearance of GAN faces could account for why participants tended to rate them as more trustworthy than real faces. Despite some empirical evidence to support the notion that increased averageness in AI faces contributes to increased realism, what makes AI faces more trustworthy remains open for discussion.

Previous research has emphasized the highly realistic nature of GAN-generated images, yet technology continues to evolve—namely, with the recent inception of diffusion models (DMs) (Ho, Jain, & Abbeel, 2020). DMs leverage high-dimensional vector spaces that encode both images and text prompts. This architecture allows the models to combine trained semantic representations into realistic depictions, even for concepts not seen before. These models are also highly accessible, allowing anyone to easily create high-quality, diverse images via simple text prompts. The sophistication of DMs increases concerns about people's ability to sort real faces from fake ones and, in turn, the extent to which DMs will contribute to the potential for misuse and a general erosion of trust (Westling, 2019).

As AI-generated images become more sophisticated and more accessible, as a society, we are increasingly exposed to artificially generated faces—often in nefarious and exploitative scenarios, such as political disinformation, financial and identity fraud, and catfishing. Hence, to understand how inferences such as trust are made, it is vital to better understand how these faces are processed and how they tap into people's preexisting biases and stereotypes. To date, however, there is little research examining whether DM-generated faces are distinguishable from real ones and no research examining human trust in these faces. A recent study, centering on computational detection of AI-generated images, preliminarily examined participants' ($N = 50$) ability to distinguish between real images and images synthesized by Midjourney (a DM) (Lu et al., 2024). When presented with 100 real and AI-generated images selected at random across eight categories (e.g., people, landscapes, objects), participants showed a limited ability (61.3% accuracy, chance = 50%) to identify whether the images were real or AI generated. Although assessing the realism of AI-generated material across various categories is interesting, synthetic facial images currently pose a unique and larger threat, given their applicability to undermine trust and damage personal identity. For

example, synthetic faces are commonly used in fake social media profiles that support fraudulent activity or to encourage societal discord. Signicat's (2024) report shows deepfake scam attempts related to identity fraud have increased by 2,137% in the past 3 years. Given the potential for misuse and harm, there is a clear need for research to establish the level of realism and trust elicited by synthetic faces generated with newer AI models (Signicat, 2024).

To date, no research has compared the trustworthiness and realism of GAN- and DM-generated faces in a highly controlled experimental study using diverse (across race, gender, and age) yet balanced and perceptually matched stimuli. In this study, we hypothesized that we would replicate the results from Nightingale and Farid's (2022) study and therefore predicted that the GAN-synthesized faces would be rated as more trustworthy than real faces. In terms of accuracy, we hypothesized that average accuracy, collapsed across all the face types, would be close to chance performance (50%). Furthermore, we hypothesized that average accuracy would be higher for real faces compared to synthetic faces and that people would perform similarly in accurately classifying GAN and DM faces. That said, as DMs are a newer architecture, we anticipated some early-stage limitations. For example, these types of models usually experience issues that need rectifying before they become stable, such as resolution mismatch, and for diffusion models in particular, inconsistencies between text prompts and synthesized images—therefore, it remained possible that DM faces might have more obvious synthesis artifacts than GAN faces, thus making them slightly easier to correctly classify as AI. Due to previous literature showing that White faces are overrepresented within some AI training datasets (Muñoz, Zannone, Mohammed, & Koshiyama, 2023), we expected White faces to be harder to correctly identify as AI compared with Black, East Asian, and South Asian faces. The overall aim of this research is to inform how trustworthy real and AI faces (GAN and DM synthesized) are and whether humans can distinguish between these.

Methods

Stimuli

We used the 400 synthetic and 400 real faces in Nightingale and Farid (2022). The 400 synthetic faces were generated using the StyleGAN2 model. To heighten ecological validity, faces were generated to be diverse across age, race, and gender. For each GAN-synthesized face, a matching real face was selected from the Flickr-Faces-HQ Dataset (FFHQ) dataset. The process for perceptually matching a real face with a synthetic face was conducted in Nightingale and Farid (2022), and we used the face pairings resulting from their work. These pairings were created in two main steps. First, a low-dimensional perceptually meaningful representation of all faces was extracted using a standard convolutional neural network descriptor based on VGG-Face (Parkhi, Vedaldi, & Zisserman, 2015). Second, the extracted representation of each synthetic face was compared with the extracted representation of all real faces in the FFHQ dataset to find the most similar real face (for more details, see Nightingale & Farid, 2022). Through this process the real images were perceptually matched to the synthetic faces. We generated 400 DM faces using Adobe Firefly's Image 2 model (Adobe, 2025). Each real face was inserted into the system to generate a matching face with respect to race and gender, as well as general appearance. A text prompt was also included, which consisted of information regarding race, gender, age, hairstyle, hair color, eye color, expression, accessories, and the statement "full head and shoulders in image" was added to ensure consistency. The stimulus set contained 1,200 images (400 real, 400 GAN, 400 DM), each with a resolution of 1,024 × 1,024 pixels. The 1,200 images were divided into 50 blocks, each containing eight real images, eight GAN-synthesized images, and eight diffusion-synthesized images, and they were balanced across race and gender. We ensured that image backgrounds remained plain and avoided using images with obvious rendering artifacts to ensure consistency with synthetic media likely to be used for nefarious purposes in real-world situations.

Experiment 1

A total of 169 participants were recruited who self-identified their race as White ($n = 74$); Black/African American ($n = 56$); East Asian ($n = 2$); Hispanic, Latino, or Spanish ($n = 13$); Middle Eastern/North African ($n = 3$); South Asian ($n = 5$); or other/prefer not to say ($n = 16$). The sample consisted of 114 women and 55 men; the mean age was 33 years ($SD = 10.87$) and ranged from 18 to 71 years. Participants were recruited via Prolific and took part in an online survey via Qualtrics. Participants were paid £3.50, and to increase task motivation, a bonus of £2.00 was awarded to participants who scored in the top 10% accuracy ratings. They were informed that failing to respond correctly to at least five of six attention checks would result in nonpayment and data exclusion—two participants were removed for failing to meet these attention checks. To ensure participants understood the task, they were first shown four real faces and informed that these were real people, and they were then shown four faces synthesized by StyleGAN and told that these people do not exist. Finally, they saw four obviously synthesized

faces with artifacts, which they were told would be similar to the faces used in the attention check trials. They then completed three practice trials to ensure they understood their task. This included one real image, one AI-synthesized image, and one AI-synthesized image with obvious artifacts to inform the attention check (AC) trials. Feedback was provided in practice trials but not during the actual task to prevent learning effects from inflating accuracy. Participants then viewed a random subset of 96 images one at a time (the initial 1,200 images were separated into 50 blocks containing an equal number of real, DM, and GAN faces). Participants were shown a total of four blocks in a randomized and counterbalanced order and, for each image, specified if the face was real or synthetic. They had unlimited time to respond ($M = 33.43$, $SD =$ 15.94 minutes).

Experiment 2

Undergraduate psychology students ($N = 87$) at Lancaster University were recruited via SONA. Participants self-identified their race as White ($n = 70$), Black/African American ($n = 1$), East Asian ($n = 5$), South Asian ($n = 6$), or other/prefer not to say ($n = 5$). The sample consisted of 77 women, 9 men, and 1 other; the mean age was 18.5 years ($SD = 0.71$). Participants' ages ranged from 18 to 21 years. Participants completed an online survey via Qualtrics and were rewarded two course credits. The students viewed 96 randomized faces (an equal number of real, DM, and GAN faces were shown in a randomized order), and participants had unlimited time to rate the trustworthiness of each face on a scale of 1 (very untrustworthy) to 7 (very trustworthy). Six attention checks were included in which a numerical rating was overlaid on the face, and participants were asked to select that number on the trustworthiness scale; three participants were removed for failing to meet these attention checks. Two participants remained inactive on the final page of the study, which substantially skewed the average completion time (104.62 minutes), but this did not compromise their data quality. The average completion time without their responses was 19.12 minutes ($SD =$ 8.39).

In both experiments, all the participants provided informed consent before beginning the study and were informed that their anonymized data would be published in an open access repository. All experiments were performed in accordance with the named guidelines and regulations. Both experiments were approved by Lancaster University's FST Research Ethics Committee (FST-2023-4010-RECR-1, FST-2024-4010-SA-1) and preregistered on the Open Science Framework; please find the URLs to the preregistrations detailing power analyses in the additional information section.

Results

Two experiments investigated the realism and trustworthiness of real, GAN-generated, and DM-generated faces. Experiment 1 explores the photo-realism of AI-generated faces. Experiment 2 provides insights into face trustworthiness.

Experiment 1

Each participant was sequentially shown a random subset of 96 faces from a pool of 1,200 faces. For each face, participants judged if it was real or AI synthesized (they were not informed of the difference between GAN- and DM-generated faces). Average accuracy was only slightly above chance at 58.4%, 95% CI [57.7%, 59.2%] ($d' = 0.55$, 95% CI [0.46, 0.64]), with a bias to respond real, $c = –0.14$, 95% CI [–0.20, –0.09]. To deal with extreme values as well as an unequal number of signal and noise trials, in our signal detection analysis, we used a method proposed by Stanislaw and Todorov (1999). This method involves an alteration to the loglinear method, whereby we added 0.33 to the hits and 0.66 to the false alarms (to account for an unequal base rate of real and AI-synthesized faces, one third of faces were real and two thirds were AI). Considering two thirds of the images were synthetic, this bias to respond that the faces were real highlights the realism of these AI-generated images. Perhaps participants tended to assume the images were real unless they had strong evidence to suggest otherwise, and in this case, the images did not have any major flaws. Participants were better at correctly classifying real (65.2%, 95% CI [63.9%, 66.4%]) and DM faces (62.1%, 95% CI [60.8%, 63.4%]) compared to GAN faces (48.0%, 95% CI [46.7%, 49.3%]), suggesting that GAN faces demonstrate greater realism. Figure 1 shows a selection of faces with their classification accuracies.

Differences in accuracy across race showed that White, East Asian, and South Asian faces were classified close to chance performance at 52.9% (95% CI [51.4%, 54.5%]), 57.0% (95% CI [55.4%, 58.5%]), and 59.6% (95% CI [58.0%, 61.1%]), while Black faces were classified at above-chance performance, 64.3% (95% CI [62.8%, 65.8%]). We hypothesize that this result may be due to Black faces being underrepresented in the model training data. The lack of racial diversity is an issue widely cited in the literature (Maluleke et al., 2022).

A generalized linear mixed-effects model was fitted to estimate accuracy from the effects of stimulus face type and race, as well as their interaction, while

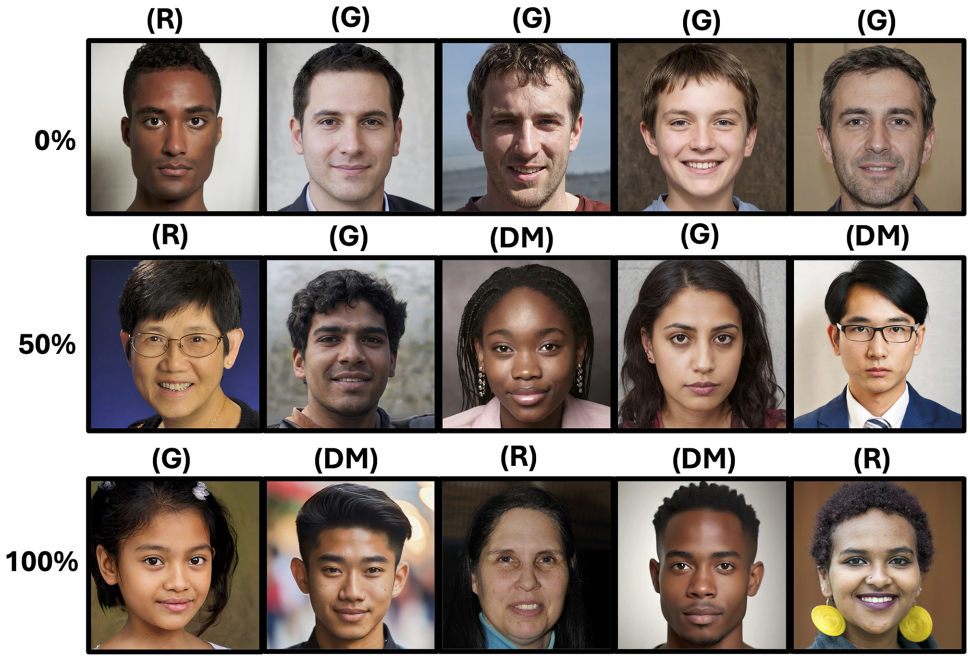

Figure 1. Representative examples of real (R), GAN (G), and diffusion (DM) faces, classified with an accuracy of 0% (top), 50% (middle), and 100% (bottom). The real faces are taken from the Flickr-Faces-HQ Dataset; the full dataset is available under the Creative Commons BY-NC-SA 4.0 license by NVIDIA Corporation.

modeling participant and item-level variability. The model included random intercepts for participants and random slopes of face type by participant. All other effects were fixed (face type, ethnicity, and their interaction). Akaike information criterion (AIC) model selection revealed that the maximal model (main effects and interaction) provided a better fit to the data than both the main effects–only model and intercept-only model. Likelihood ratio tests confirmed that the improvement in model fit between the main effects

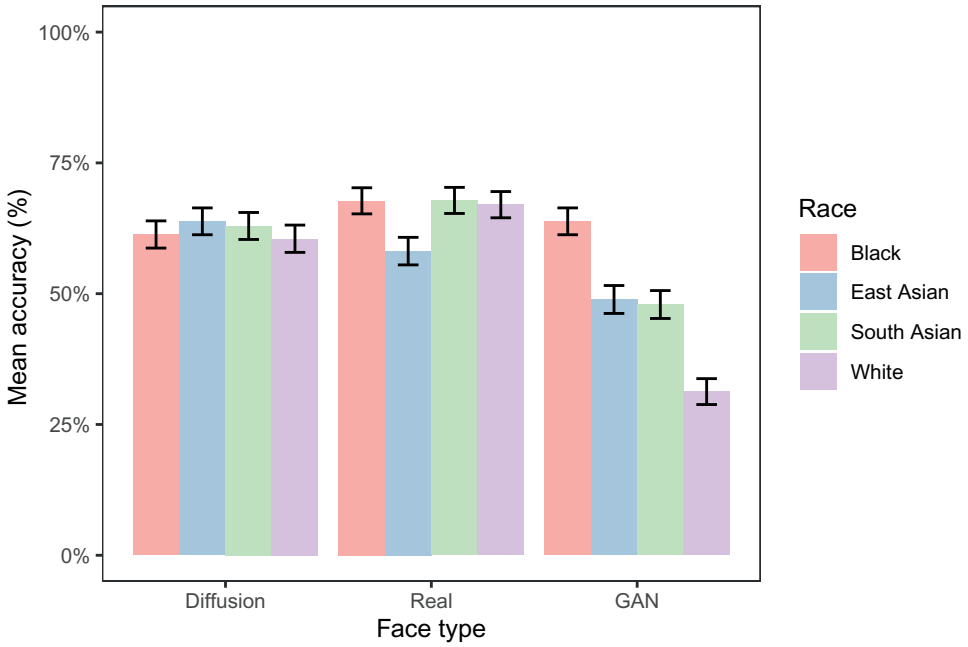

Figure 2. A bar plot showing the interaction between face type and race on mean percentage accuracy. The interaction is driven by a low mean accuracy rating for GAN faces, particularly White GAN faces, suggesting these faces were the most difficult to classify correctly. The error bars display 95% confidence intervals.

model and the interactions model was significant, $\chi 2(18) = 259.09$, $p < 0.001$. The significant interaction (see Supplementary Table S1 for the comprehensive model outputs) suggests the effect of face type is not consistent across all races. In comparison to White GAN faces:

- DM White faces were 4.05 times more likely to be correctly classified, 95% CI [3.32, 4.94].
- Real White faces were 5.55 times more likely to be correctly classified, 95% CI [4.30, 7.16].

This suggests that White GAN faces were the hardest to classify, and this difficulty drives the interaction shown in Figure 2.

Experiment 2

Participants were sequentially shown a different subset of 96 faces selected from a pool of 1,200 faces. Participants rated each face on a scale from 1 (very untrustworthy) to 7 (very trustworthy). A one-way, repeated-measures analysis of variance revealed a significant effect of face type on trustworthiness ratings, $F(2, 172) = 131.3$, $p < 0.001$, $\eta^2 G = 0.22$), highlighting that trustworthiness scores differed across each face type. Post hoc tests with Bonferroni comparisons applied revealed that real faces proved the least trustworthy, with a trust rating of 4.03, 95% CI [3.90, 4.15], $SD = 1.59$. As predicted from previous literature (Nightingale & Farid, 2022), GAN faces were more trustworthy than real faces, with a trust rating of 4.36, 95% CI [4.26, 4.47], $SD = 1.52$, $t(86) = 8.61$, $p < 0.001$, $d = 0.64$. DM faces proved the most trustworthy, with an average of 4.70, 95% CI [4.59, 4.81], $SD = 1.49$. They were more trustworthy than both real faces, $t(86) = 14.7$, $p < 0.001$, $d = 1.26$, and GAN faces, $t(86) = 8.52$, $p < 0.001$, $d = 0.68$. All of the face types (real, GAN, and diffusion) received a minimum trust rating of 1 and a maximum trust rating of 7. In addition, the percentage of faces that were manually determined as visibly smiling was similar across groups (real faces mean = 61.75%, GAN faces mean = 65.5%, diffusion faces mean = 65.0%).

Discussion

Although DM faces were easier to identify as synthetic than GAN faces, most AI-generated faces were sufficient to fool most people much of the time. In addition, DM faces were rated as more trustworthy than real or GAN faces. One possible explanation for the realism finding is that Firefly has not yet reached the sophistication of StyleGAN for producing intricate details of human faces (e.g., skin texture), perhaps due to the different model architectures (Karn, Kumar, Kushwaha, & Katarya, 2023). The GAN faces, for example, were generated by a model trained to create only faces, whereas DMs are trained to create images across any category, ranging from faces to animals, objects, landscapes, and so on. In addition, GAN models use the truncation trick, a technique to amplify the quality of the images produced. Essentially, the generator is fed with less random noise; therefore, the sampling data are skewed toward the mean of the distribution.

In line with our hypothesis and previous research (Miller et al., 2023; Nightingale & Farid, 2022), the StyleGAN faces were rated as more trustworthy than real faces. We hypothesize that this finding may be due to the typicality of the faces produced by StyleGAN, a model trained predominantly on real faces from the FFHQ. The model therefore tends to be biased toward generating faces close to the average of those faces, which has subsequently been supported by human perceptual research, where raters judged GAN faces to appear as more average than real faces (Miller et al., 2023). Diffusion models are also biased toward producing faces close to the mean of the training data distribution. Adobe Firefly is trained on a much larger pool of Adobe Stock images; therefore, rare features are lost during the averaging process. Drawing on face space theory (Valentine, 1991), it is possible to predict why higher averageness might lead to higher trust ratings for AI faces than for real faces. Faces seen more often cluster together and establish stronger representations, while faces we experience less frequently are sparsely distributed in our representation; hence, average faces may appear familiar, leading to increased trust in them. From this and prior research (Sofer et al., 2015), we might conclude that facial averageness contributes to high levels of trust.

DM faces were rated as less realistic than GAN faces but more trustworthy than real or GAN faces. This presents an interesting paradox and suggests that realism judgments—and more social evaluations, such as trustworthiness—might be driven by two distinct psychological mechanisms. Therefore, it seems that rating a face as trustworthy is not solely dependent on the level of realism and instead might depend on the perception of certain social traits, for example, attractiveness, valence, or simply facial expressions. Ultimately, previous suggestions about the averageness of AI-generated faces accounting for both high levels of photo-realism and trustworthiness might not be the whole story.

Multiple explanations may account for why DM faces were rated as most trustworthy. We know that Firefly was trained on Adobe Stock images (Adobe, 2025), which tend to include people in more professional

attire/poses. As such, one hypothesis is that this training set might have biased the resulting synthesized images toward a higher level of professionalism, which in turn might induce higher ratings of trust. Connections have been made between valence and trust (Oosterhof & Todorov, 2008); one predominant visible cue that can indicate positive valence is a smile. Previous literature highlights the association between positive valence and trust (Galinsky et al., 2020), leading us to question if variations in expression across the faces could explain this finding, for example, whether DM faces were rated as the most trustworthy because of a higher percentage of smiling faces. However, the percentage of smiling faces across the stimulus pool proved highly similar. Therefore, we do not believe that positive valence alone accounts for the differences in trust reported here.

Another possibility is that the Firefly training images might include more faces with aesthetic/stylistic filters that aim to "beautify"; if so, these filtered effects are likely reinforced by the DM. Such filtered faces may tune into specific attributes that heighten trust, for example, attractiveness. Previous research provides evidence of a correlation between attractiveness and trust (Wilson & Eckel, 2006), whereby attractive faces have been linked to higher trust ratings (Shinners, 2009). We hypothesize that the inclusion of heavily filtered faces in Adobe Firefly's training data might have contributed to DM faces being perceived as more attractive and, as such, influenced ratings of trust, although the relationship between attractiveness and trust is not entirely clear (Sofer et al., 2015). In addition, today's society is highly exposed to photos of filtered faces—much more so than natural ones—which could in turn make such faces feel more familiar. Familiar faces tend to be rated as more trustworthy; in one study, for example, participants rated same-sex faces (computationally determined) to have higher self-resemblance as more trustworthy than faces with lower self-resemblance (Nakano & Yamamoto, 2022). Further research could use computational measures of expression, valence, and perceived attractiveness ratings to disentangle which attributes drive the observed differences in trustworthiness across AI-generated faces. It currently remains unclear which or what combination of these factors is driving the increased trustworthy appearance in AI-generated faces.

In the current study, participants also showed a bias to respond "real," despite the fact that two thirds of the faces they saw were AI generated. Therefore, when making realism judgments, it is possible that people default to assuming authenticity in the absence of anything appearing obviously "off." Perhaps the fast-paced technological changes in society, paired with information overload, encourage humans to make rapid decisions about others and the world around us—and the default position might be to accept what is seen as being real.

In addition to the wide range of nefarious uses of AI-generated faces (e.g., identity fraud and catfishing), another concern is the potential for perpetuating racial biases. Consistent with previous literature, we find a potential racial bias within the StyleGAN model (Cheong et al., 2024). Of the four races examined, White GAN faces proved the most realistic, likely due to differences in training data: StyleGAN is trained on the FFHQ dataset, often criticized for lacking racial diversity, and Firefly is trained on open-source images from Adobe stock, which may encompass a wider pool of more diverse races. Biased image generation by these models is concerning, as it can perpetuate and reinforce society's existing racial and cultural biases. For example, when asked to generate a "Latina woman," Stable Diffusion generated a sexually explicit image contributing to gender and racial stereotypes (Post, 2023). Of course, it is possible that racial differences between the participants and stimuli may have influenced the results in this study; therefore, future research should prioritize controlling for these effects.

Given limits in human ability to distinguish real from fake images, attempts have been made to improve perceptual performance, although most have had limited success. Training to raise awareness of synthesis artifacts led to only small improvements in accuracy (Nightingale & Farid, 2022), and media literacy training reduced belief in AI-generated images but reduced belief in real information too (Guo, Swire-Thompson, & Hu, 2025). Other attempts to combat misinformation include watermarking (Nadimpalli & Rattani, 2024); nevertheless, ensuring robustness to tampering is difficult. Platforms such as the Content Provenance and Authenticity (C2PA) (Rosenthol, 2022) and the Content Authenticity Initiative (CAI) can help by encouraging technical standards that ensure images retain provenance information. We encourage transparency and urge large corporations to adopt this standard.

Conclusions

To conclude, AI faces are highly realistic and trustworthy, even more so than real faces. We compared GAN and DM-synthesized faces, highlighting how technological advancements and sophistication in image generation bring new levels of uncertainty and the ability to deceive. More work is required to underpin how these faces are processed by the human eye. The juxtaposition between realism and trust ratings requires further attention; further work should investigate the differing attributes across real

and AI-synthesized faces to provide insight into the informants of realism and trust. Democratized access to generative AI further propels risks for individuals, societies, and democracy—there is an urgent need to address the ever-developing threat landscape facilitated by the frontiers of technology.

Keywords: deep fakes, face perception, synthetic media

Acknowledgments

The authors thank Matthew Ivory and Matthew Asher for their contributions.

Funded by the Centre for Research and Evidence on Security Threats (ESRC Award: ES/V002775/1), which is funded in part by the U.K. Home Office and security and intelligence agencies (see the public grant decision: https://gtr.ukri.org/projects?ref=ES%2FV002775%2F1). The funding arrangements required this article to be reviewed to ensure that its contents did not violate the Official Secrets Act or disclose sensitive, classified, and/or personal information.

Supported by Security Lancaster. All authors have been involved in the design of the experiments and drafting of the manuscript and have read and approved the final version. The real images used in this study were taken from Nightingale and Farid (2022) and originally sourced from the Flickr-Faces-HQ Dataset (FFHQ). These individual images were published under a range of licenses (Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license), which permit the use, distribution, and adaptation for noncommercial purposes. The GAN-synthesized faces used in the study are derived from Nightingale and Farid (2022) and are open source. The faces generated using Adobe Firefly were based on the FFHQ faces; we used publicly available faces and adhered to the licensing terms. Faces produced by the Firefly model are not identifiable individuals.

Data availability: Preregistrations are available https://osf.io/gd2sp; https://osf.io/mjp4n and here. All materials and data underlying this study can be found https://osf.io/ygeau; https://osf.io/j7d6v.

Commercial relationships: Hany Farid is affiliated with C2PA and CAI initiatives.
Corresponding author: Alexis A. McGuire.
Email: l.mcguire@lancaster.ac.uk.
Address: Psychology Department, Lancaster University, Lancaster, Lancashire LA1 4YW, UK.

References

Adobe. (2025). Adobe Firefly, https://www.adobe.com/uk/products/firefly.html.

Atkinson, A. P., & Adolphs, R. (2011). The neuropsychology of face perception: Beyond simple dissociations and functional selectivity. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*, 1726–1738.

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity, 9*, tyad011.

Cheong, M., Abedin, E., Ferreira, M., Reimann, R., Chalson, S., Robinson, P., . . . Klein, C. (2024). Investigating gender and racial biases in dall-e mini images. *ACM Journal on Responsible Computing, 1*(2), 1–20.

Galinsky, D. F., Erol, E., Atanasova, K., Bohus, M., Krause-Utz, A., & Lis, S. (2020). Do I trust you when you smile? Effects of sex and emotional expression on facial trustworthiness appraisal. *PLoS One, 15*(12), e0243230.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*, 2672–2680.

Guo, S., Swire-Thompson, B., & Hu, X. (2025). Specific media literacy tips improve AI-generated visual misinformation discernment. *Cognitive Research: Principles and Implications, 10*(1), 38.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Karn, A., Kumar, S., Kushwaha, S. K., & Katarya, R. (2023). Image synthesis using GANs and diffusion models. In *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)* (Vol. *1*, pp. 1–6). IEEE.

Lago, F., Pasquini, C., Böhme, R., Dumont, H., Goffaux, V., & Boato, G. (2021). More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine, 39*(1), 109–116.

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., . . . Ouyang, W. (2024). Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems, 36*, 25435–25447.

Maluleke, V. H., Thakkar, N., Brooks, T., Weber, E., Darrell, T., Efros, A. A., . . . Guillory, D. (2022). Studying bias in GANs through the lens of race. In *European Conference on Computer Vision* (pp. 344–360). Cham: Springer Nature Switzerland.

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A., Krumhuber, E. G., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science, 34*(12), 1390–1403.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*, 98–100.

Muñoz, C., Zannone, S., Mohammed, U., & Koshiyama, A. (2023). *Uncovering bias in face generation models*. arXiv preprint arXiv:2302.11562.

Nadimpalli, A. V., & Rattani, A. (2024). Proactive deepfake detection using GAN-based visible watermarking. *ACM Transactions on Multimedia Computing, Communications, and Applications, 20*, 1–27.

Nakano, T., & Yamamoto, T. (2022). You trust a face like yours. *Humanities and Social Sciences Communications, 9*, 1–6.

Nightingale, S. J., & Farid, H. (2022). Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America, 119*, e2120481119.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80*, 381.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092.

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015.* British Machine Vision Association.

Post, T. W. (2023). *This is how AI image generators see the world*, https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/.

Rosenthol, L. (2022). C2PA: The World's first industry standard for content provenance (Conference presentation). In A. G. Tescher & T. Ebrahimi (Eds.), *Applications of Digital Image Processing XLV (Vol. 12226, Article 122260P)*. San Diego, CA, USA: SPIE.

Shinners, E. (2009). Effects of the "what is beautiful is good" stereotype on perceived trustworthiness. *Undergraduate Research & Creativity*, *12*, 1–5.

Signicat. (2024). *The battle against AI-driven identity fraud*, https://www.signicat.com/the-battle-against-ai-driven-identity-fraud.

Sofer, C., Dotsch, R., Wigboldus, D. H., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science, 26*(1), 39–47.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137–149.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*, 455–460.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43*, 161–204.

Westling, J. (2019). Are deep fakes a shallow concern? A critical analysis of the likely societal reaction to deep fakes. In *A Critical Analysis of the Likely Societal Reaction to Deep Fakes (July 24, 2019). TPRC47: The 47th Research Conference on Communication,* Information and Internet Policy.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*, 592–598.

Wilson, R. K., & Eckel, C.C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly, 59*(2), 189–202.