

AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones



Elizabeth J. Miller¹, Ben A. Steward¹, Zak Witkower²,
Clare A. M. Sutherland^{3,4}, Eva G. Krumhuber⁵, and
Amy Dawel¹

¹School of Medicine and Psychology, Australian National University; ²Department of Psychology, University of Toronto; ³School of Psychology, King's College, University of Aberdeen; ⁴School of Psychological Science, University of Western Australia; and ⁵Department of Experimental Psychology, University College London

Psychological Science
1–14

© The Author(s) 2023



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09567976231207095

www.psychologicalscience.org/PS



Abstract

Recent evidence shows that AI-generated faces are now indistinguishable from human faces. However, algorithms are trained disproportionately on White faces, and thus White AI faces may appear especially realistic. In Experiment 1 ($N = 124$ adults), alongside our reanalysis of previously published data, we showed that White AI faces are judged as human more often than actual human faces—a phenomenon we term *AI hyperrealism*. Paradoxically, people who made the most errors in this task were the most confident (a *Dunning-Kruger effect*). In Experiment 2 ($N = 610$ adults), we used face-space theory and participant qualitative reports to identify key facial attributes that distinguish AI from human faces but were misinterpreted by participants, leading to AI hyperrealism. However, the attributes permitted high accuracy using machine learning. These findings illustrate how psychological theory can inform understanding of AI outputs and provide direction for debiasing AI algorithms, thereby promoting the ethical use of AI.

Keywords

artificial intelligence, face perception, face-space theory, generative adversarial network, StyleGAN2

Received 4/25/23; Revision accepted 9/19/23

The dawn of the artificial intelligence (AI) revolution has marked an unprecedented societal shift (Xie, 2023). Prominent in this shift is the generation of realistic humanlike AI faces, twinned with public concern that AI might distort the perception of truth (Devlin, 2023). AI-generated faces are now widely available (e.g., this-person-does-not-exist.com) and are being used for both prosocial and nefarious purposes, from finding missing children (Chandaliya & Nain, 2022) to transmitting political misinformation via fake social media accounts (e.g., Hatmaker, 2020). AI faces are now so realistic that people often fail to detect they are not human (e.g., Nightingale & Farid, 2022). However, because this technology has advanced so rapidly (Hao, 2021), there have been few empirical tests of this ability. Here we argue that AI faces are not just indistinguishable from human faces but that, in fact, they may be perceived as more “human” than real human faces. We term this striking and counterintuitive phenomenon *hyperrealism*. The

present research aimed to test for and explain AI hyperrealism, investigate whether people have insight into their AI detection errors, and discover visual attributes that can be used to reveal AI imposters.

Psychology offers decades of theoretical and empirical work with potential to explain AI hyperrealism. For example, the influential *face-space theory* (Fig. 1; Valentine, 1991; Valentine et al., 2016) proposes a hypothetical multidimensional space in which faces are coded along unspecified dimensions according to how much they differ from an average face located at the center. Human faces are supposed to be normally distributed within this space in such a way that more average features (for all dimensions) are statistically

Corresponding Author:

Amy Dawel, Australian National University, School of Medicine and Psychology

Email: amy.dawel@anu.edu.au

overrepresented. This bias toward average features that generative algorithms are trained on (e.g., StyleGAN2 for faces; Karras et al., 2020, 2021) may be further exaggerated in the AI faces they generate, as these algorithms are biased toward the most common statistical properties of their training data (Grossman et al., 2019). Although the specific dimensions of face-space are unknown, it is possible to measure the relative location of faces indirectly via the emergent perceptual attributes of face-space, such as facial averageness. Thus, we hypothesized that StyleGAN2-generated faces would embody the attributes of average faces to a greater extent than real human faces.

A psychological analysis of AI representativeness can also help with understanding a puzzle arising from the handful of studies that have investigated people's ability to detect AI faces: Although one recent study found that people were unable to distinguish AI from human faces (Nightingale & Farid, 2022), two others go further to suggest that people may overidentify AI faces as human (Shen et al., 2021; Tucciarelli et al., 2022). How can we explain this puzzle? All three studies used the

Statement of Relevance

Artificial intelligence, or AI, can now generate faces that are indistinguishable from human faces. However, AI algorithms tend to be trained using a disproportionate number of White faces. As a result, AI faces may appear especially realistic when they are White. Here, we show that White (but not non-White) AI faces are, remarkably, judged as human more often than pictures of actual humans. We pinpoint the perceptual qualities of faces that contribute to this hyperrealism phenomenon, including facial proportions, familiarity, and memorability. Problematically, the people who were most likely to be fooled by AI faces were the least likely to detect that they were being fooled. Our results explain why AI hyperrealism occurs and show that not all AI faces appear equally realistic, with implications for proliferating social bias and for public misidentification of AI.

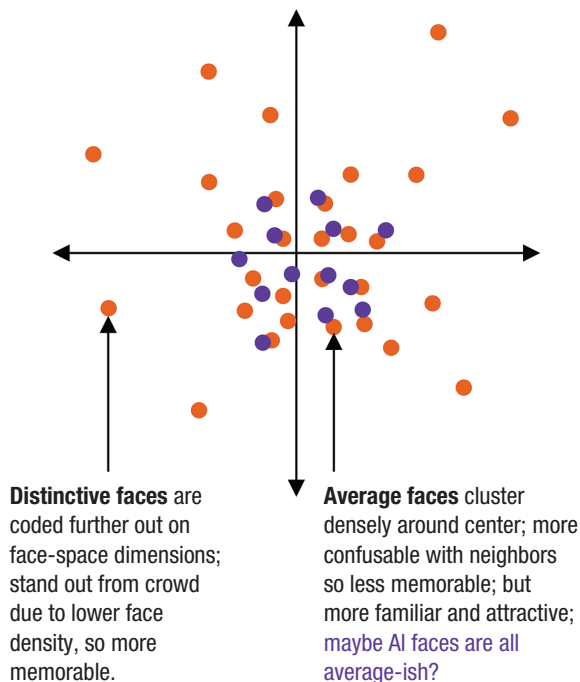


Fig. 1. Schematic illustration of face-space theory: A potential explanation for AI hyperrealism. Orange dots show sample distribution of human faces; purple dots show hypothesized distribution of AI faces. We focus on relevant abstract principles of face-space theory (e.g., relating to single images of faces in human perception). For more nuanced discussions, see Burton et al. (2016), O'Toole et al. (2018), and Valentine et al. (2016). For psychophysics-related work, see Abudarham and Yovel (2016) and Rhodes and Jeffery (2006).

StyleGAN2 algorithm but varied in the race of the faces they tested. These demographic differentials are critical because StyleGAN2 was trained on primarily White faces (~69% White, ~31% for all other races combined; see Supplemental File S1 in the Supplemental Material available online), potentially biasing the algorithm toward the statistical regularities of White faces. This bias may lead to White AI faces that appear especially average (indicated in Fig. 1) and therefore, potentially, especially realistic. Consistent with this theory, Shen et al. (2021) and Tucciarelli et al. (2022) found preliminary evidence of AI hyperrealism to the extent they tested White faces, although Tucciarelli et al.'s (2022) stimuli were also preselected to be particularly realistic, biasing them toward this finding. Intriguingly, Nightingale and Farid (2022) also reported more errors for White than non-White AI face detection. However, they did not pursue this question further. If AI faces do appear more realistic for White faces than other groups, their use will confound perceptions of race with perceptions of being "human." Thus, the use of popular StyleGAN2 faces may risk misleading scientific conclusions (Dawel et al., 2022) and may even perpetuate social biases in real-world outcomes, from influencing elections to finding missing children (Chandaliya & Nain, 2022; Hatmaker, 2020).

The Present Research

Here we aimed to investigate the potential for AI hyperrealism and provide the first test of whether people

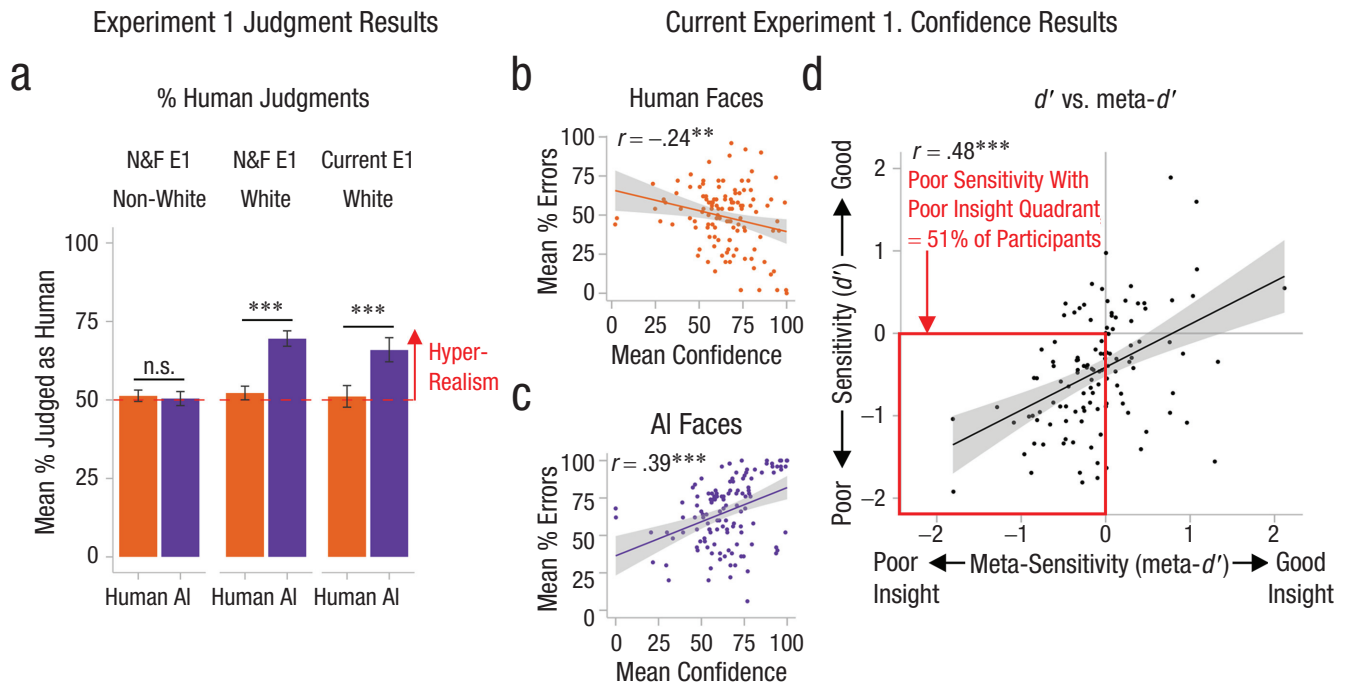


Fig. 2. Reanalysis of data from Experiment 1 of Nightingale and Farid (2022) and results for current Experiment 1. Error bars represent 95% confidence intervals. N&F E1 = data from Nightingale and Farid (2022), Experiment 1; n.s. = nonsignificant.

$**p < .01$. $***p < .001$.

have insight into their AI detection errors. If people mistake AI faces as human but have low confidence in their judgment, they may respond more cautiously (e.g., investigating an online profile). However, if they are convinced their judgment is correct, their errors may be more consequential (e.g., falling for a fraudulent profile). Although Tucciarelli et al. (2022) found confidence was higher for judgments of AI than for human faces overall, it is currently unknown whether people are aware of their AI detection errors. Errors are associated with lower confidence for other face judgments (e.g., face identity recognition, Palermo et al., 2017; eyewitness identification, Wixted & Wells, 2017). Thus, we predicted that poorer AI detection would be associated with lower confidence.

We also aimed to identify visual attributes that distinguish AI from human faces and address the critical unanswered question of why people fail to detect AI faces. Our theorizing suggests that the emergent perceptual attributes of face-space—such as facial averageness, memorability, attractiveness, and familiarity—may play a role, given their importance for human face perception (Valentine et al., 2016; Vokey & Read, 1992). Because little is known about which cues people use for AI detection, we augmented this theoretical perspective with a data-driven approach by asking participants what information they used to guide their judgments.

Reanalysis of Nightingale and Farid (2022)

We started with a proof-of-principle by reanalyzing data from a prominent recent study that included information about face race (Nightingale & Farid, 2022, Experiment 1) to investigate the potential for AI hyperrealism. Analyses showed clear evidence of AI hyperrealism for White faces, but not for non-White faces. Figure 2a shows that White AI faces were judged as human significantly more often than White human faces, $M_{\text{White-AI}} = 69.5\%$ versus $M_{\text{White-human}} = 52.2\%$, $t(314) = 13.25$, $p < .001$, $d = 0.75$, 95% confidence interval (CI) = [0.62, 0.87], and significantly more often than chance (= 50% in the two-alternative forced choice task), $t(314) = 16.01$, $p < .001$, $d = 0.90$, 95% CI = [0.77, 1.03]. In contrast, non-White AI faces (left side of Fig. 2a) were judged as human at around chance levels, $M_{\text{non-White-AI}} = 50.5\%$, $t(314) = 0.41$, $p = .682$, $d = 0.02$, 95% CI = [-0.09, 0.13], which did not differ significantly from how often non-White human faces were judged to be human, versus $M_{\text{non-White-human}} = 51.3\%$, $t(314) = 0.74$, $p = .461$, $d = 0.04$, 95% CI = [-0.07, 0.15]. Unusually, d' —a measure of people's ability to discriminate between AI and human faces that is unaffected by response bias—was also significantly negative for White faces, $M = -.59$, $t(314) = 13.17$, $p < .001$, $d = 0.74$, 95% CI = [0.62, 0.87]. The d'

result indicates that participants did discriminate between White AI and human faces, but in the wrong direction, providing clear evidence of AI hyperrealism for White faces.

Experiment 1

To investigate whether people have insight into their AI hyperrealism errors and uncover what causes this somewhat counterintuitive phenomenon, we asked a new set of participants to report how confident they felt, and what information they used, when attempting to distinguish AI from human faces. Focusing our new empirical work on the White AI faces from Nightingale and Farid (2022) enabled us to test the robustness of AI hyperrealism with a new set of participants.

Open practices statement

We report all measures and exclusions (see Supplemental File S2), along with power analyses justifying our sample sizes (see Supplemental File S3). Data, analysis scripts, and materials are available on the Open Science Framework at osf.io/sz2fe/. Stimuli are available at osf.io/ru36d/. Data were analyzed using R version 4.2.1 (R Core Team, 2021) and JASP (JASP Team, 2023). Meta- d' was calculated in MATLAB (The MathWorks, Natick, MA; Maniscalco & Lau, 2012, 2014).

Method and participants

The final data were from 124 adults (61 men, 62 women, 1 preferred another term) recruited from Prolific (www.prolific.co). Participants were White U.S. residents, aged 18 to 50 years ($M_{\text{age}} = 34.4$ years, $SD = 8.0$ years), who had not lived outside the United States for more than 2 years before they turned 18 and who reported not having autism spectrum disorder, attention deficit disorder/attention deficit hyperactivity disorder, schizophrenia, or a major neurological condition. We recruited only White participants because of potential out-group effects in humanness ratings (McLoughlin et al., 2018) and other-race effects (McKone et al., 2019; Meissner & Brigham, 2001).¹ Participants' data were excluded if they did not complete the full study or missed > 10% of experimental trials in the AI or human face conditions; used a mobile phone (because face stimuli would not appear appropriately sized); responded incorrectly on more than one attention check question; or responded incorrectly when they were asked at the end of the study what task they had performed (see Supplemental File S2). All participants in this research provided

informed consent, and all experiments were approved by the Australian National University Human Research Ethics Committee (Protocol 2019/970).

Stimulus materials

We used the 100 AI and 100 human White faces (half male, half female) from Nightingale and Farid (2022; see osf.io/ru36d/). The AI faces were generated using StyleGAN2. The human faces were selected from the Flickr-Faces-HQ Dataset (Karras et al., 2021, used to develop StyleGAN2) to match each of the AI faces as closely as possible (e.g., same gender, posture, and expression). All stimuli had blurred or mostly plain backgrounds, and AI faces were screened to ensure they had no obvious rendering artifacts (e.g., no extra faces in background). Screening for artifacts mimics how real-world users screen AI faces, either as scientists (Peterson et al., 2022) or for public use (Satter, 2021), and therefore captures the type and range of stimuli that appear online. Participants were asked to resize their screen so that stimuli had a visual angle of 12° wide × 12° high at ~50 cm viewing distance.

Participants were assigned in counterbalanced order by gender to view either all the male or female faces (50 AI + 50 human faces = 100 trials in total) so that approximately equal numbers of men and women were assigned to each face sex. Faces were shown individually until a response was made, with order randomized across participants.

Procedure

Participants were told that they would see approximately 100 faces with the task of deciding whether each face depicts a real human or is computer-generated (AI). We defined "human" as people who exist in the real world and "computer-generated" as pictures that have been made by AI technology for generating highly realistic images of people who do not exist in the real world. After deciding whether a face was AI or human, participants rated their confidence on each trial from 0 (*not at all*) to 100 (*completely*). The AI and human response options were shown horizontally on screen with the left/right position counterbalanced across participants. In five additional trials, as an attention check, participants decided whether a face was under or over 50 years of age (see Supplemental File S4). Finally, to investigate the visual attributes used by participants to judge whether faces were AI or human, we asked participants to give open-ended responses about what information they used. At the end of the experiment, participants reported their age, gender identity, time

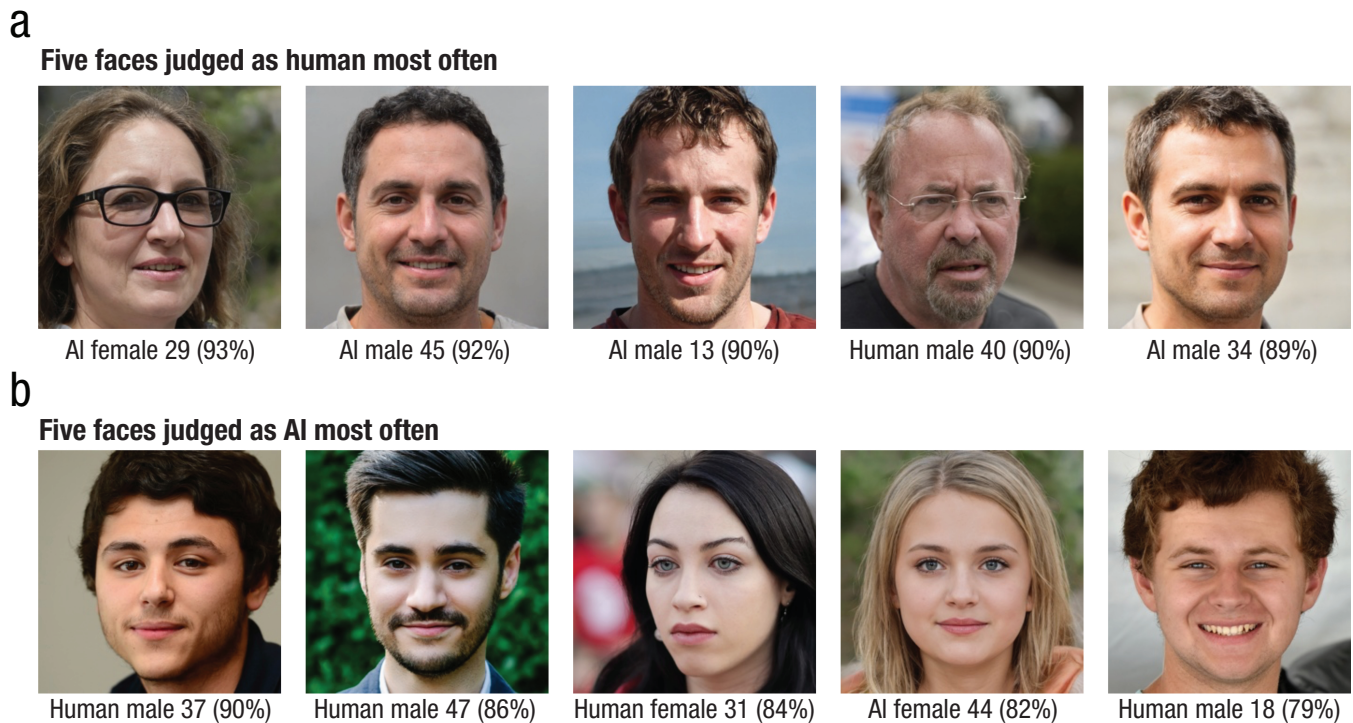


Fig. 3. Faces judged most often as (a) human and (b) AI. The stimulus type (AI or human; male or female), the stimulus ID (Nightingale & Farid, 2022), and the percentage of participants who judged the face as (a) human or (b) AI are listed below each face.

spent outside the United States, state of residence, and any clinical or neurological conditions, and they confirmed they were White.

Results

Analytic strategy

First, we calculated the percentage of stimuli judged as human, the error percentages, and the mean confidence ratings for each participant (for AI and human faces separately; Supplemental File S5). Complementary stimulus-level analyses are reported in Supplemental File S6. We also calculated participant-level signal detection measures: d' and meta- d' . Meta- d' combines confidence ratings with response correctness to measure metacognitive sensitivity—participants' insight into whether their responses are correct or incorrect (Maniscalco & Lau, 2012, 2014) (Supplemental File S7). To analyze the open-ended qualitative responses, we used data-driven (inductive) thematic analysis following the stages proposed by Braun and Clarke (2006). The first author initially read through the responses and formulated an initial thematic framework. The themes were then refined and finalized via detailed discussion with the second and last authors.

AI hyperrealism is robust

Figure 2a shows that the hyperrealism found for White AI faces in our reanalysis of Nightingale and Farid (2022) was fully replicated in our new sample, indicating that this effect is robust. White AI faces were judged as human significantly more often than White human faces, $M_{AI} = 65.9\%$ versus $M_{human} = 51.1\%$, $t(123) = 7.14$, $p < .001$, $d = 0.64$, 95% CI = [0.45, 0.83], and significantly more often than chance (50% in 2AFC task), $t(123) = 7.82$, $p < .001$, $d = 0.70$, 95% CI = [0.51, 0.90]. In contrast, performance for White human faces did not differ significantly from chance, $t(123) = 0.60$, $p = .550$, $d = 0.05$, 95% CI [-0.12, 0.23]. As in our reanalysis of Nightingale and Farid (2022), d' was significantly negative, $d' = -0.49$ (vs. 0 = no sensitivity), $t(123) = 5.20$, $p < .001$, $d = 0.68$, 95% CI = [0.49, 0.88]. Findings held for male and female faces separately (Supplemental File S8). Figure 3 shows the faces that were judged as human and AI most often: Notably, the top three most humanlike faces were actually AI-generated.

Do people have insight into their AI detection errors?

Concerningly, we found that participants who were the worst at detecting AI faces had the poorest insight

into their abilities, against our prediction from the face identification literature. However, the accuracy-confidence relationship differed by face type: Although lower error rates for classification of human faces were associated with higher confidence as predicted, $r_{\text{human}} = -.235$, 95% CI = $[-.395, -.061]$, $p = .009$ (Fig. 2b), for AI faces more errors were unexpectedly associated with higher confidence, $r_{\text{AI}} = .385$, 95% CI = $[.224, .526]$, $p < .001$ (Fig. 2c), indicating that the tendency for AI hyperrealism is exacerbated by overconfidence.

To investigate participants' insight into their performance, free from bias in confidence ratings (e.g., reporting high confidence for all judgments), we used meta- d' . Positive meta- d' values indicate participants have insight into whether their responses are correct or incorrect, whereas negative meta- d' values indicate participants have misplaced confidence in the correctness of their responses. Participants' meta- d' values in the present study were frequently negative (59% of participants), indicating poor insight. Moreover, Figure 2d shows lower insight (meta- d') was associated with poorer performance (d') on the AI versus human judgment task, $r = .479$, 95% CI = $[.330, .604]$, $p < .001$, indicating the poorest performers were the least aware of their AI detection errors. We divided Figure 2d into quadrants at $d' = 0$ and meta- $d' = 0$ to identify groups of participants with each combination of good and poor performance (d') and insight (meta- d'). Half of the participants (51%) fell into the bottom left quadrant, pairing poor performance with poor insight (versus ~23% with poor performance and good insight; ~8% with good performance and poor insight; and ~18% with good performance and good insight). These findings are incongruent with prior face perception literature (Palermo et al., 2017; Wixted & Wells, 2017) but in line with other types of judgments in which people can be highly confident but incorrect (e.g., when people are unknowingly exposed to misinformation but report it with high confidence; Flowe et al., 2019), or overestimate their competence at a task, commonly referred to as the *Dunning-Kruger effect* (Kruger & Dunning, 1999).

What visual attributes do participants report using to judge faces as AI versus human?

Figure 4 presents the qualitative coding framework capturing the attributes that participants reported using when they judged whether faces were AI or human. The size of each segment indicates the percentage of total codes captured by each theme. The framework is composed of 21 main themes with 20 subthemes (e.g., "eyes" is a subtheme of the specific facial features

theme). Responses could be coded into multiple themes, and thus each response was coded into an average of 2.29 themes. For instance, the response, "If the faces were overly symmetrical and if they [sic] eyes looked fake" was coded into the "symmetry," "eyes," and "artificial" themes. A total of 546 codes were applied to the 239 responses. Supplemental File S9 includes example quotes for each theme.

Experiment 2

The phenomenon of AI hyperrealism implies there must be some visual differences between AI and human faces, which people misinterpret. Very little is known about what these differences might be. Tucciarelli et al. (2022) found a partial negative contribution of attractiveness, which aligns with our predictions based on face-space, because faces at the core of face-space (more average faces) tend to be more attractive, all else being equal (Rhodes, 2006). Shen et al. (2021) also found that removing background scenery made AI and human faces indistinguishable; however, background information was matched for our stimuli, rendering this latter explanation unlikely here.

Thus, in Experiment 2 we investigated the capacity of 14 attributes derived from face-space and Experiment 1 qualitative reports to explain AI hyperrealism. We also tested for the first time whether human-perceivable information can be used to accurately classify AI and human faces, using machine learning. If, as we hypothesize, StyleGAN2 is biased to produce faces toward the center of face-space, AI faces should be perceived as more average, familiar, and attractive, but as less memorable than human faces.

Method

Participants

The final data were from 610 participants (290 men, 312 women, 8 preferred another term; M age = 35.3 years, $SD = 8.6$ years), recruited to rate the AI and human faces on one of 14 attributes. In contrast to Experiment 1, participants were not told AI faces were present, and we excluded those who guessed that AI faces were present ($N = 44$, 7%). Participant screening was otherwise identical to Experiment 1.

Procedure

In total, 14 attributes were rated (Table 1). In addition to the four attributes derived from face space theory (distinctiveness/averageness, memorability, familiarity,



Fig. 4. Qualitative responses from Experiment 1: percentage of codes ($N = 546$) in each theme. Subthemes are shown at the outside edge of the main theme.

attractiveness), we focused our analyses on attributes commonly mentioned in Experiment 1, resulting in nine attributes. We also included perceived age because we wanted to isolate the contributions of other related attributes, such as attractiveness and skin smoothness. Supplemental File S10 provides a detailed rationale. Each condition had five attention checks that asked for specific numeric ratings. Experimental stimuli and procedure were otherwise identical to Experiment 1.

Results

Analytic strategy

We calculated the stimulus-level mean rating for each face for each of the 14 attributes separately. Then, using our data from Experiment 1, we calculated the percentage of participants who judged each face as human. Higher percentage values indicate that more participants judged the face as human. Stimulus type (i.e., AI or human faces) was dummy-coded (0 = AI faces and 1 = human faces).

Which visual attributes contribute to faces being judged as human?

To determine what attributes made faces look real (even if they were AI-generated), we constructed a multiple linear regression model predicting the percentage of participants who judged each stimulus as human from the 14 stimulus-level attribute means.² All variables were standardized prior to model entry. The model explained the majority of observed variance (62%) in how often faces were judged as human, $R^2_{adj} = .62$, 95% CI = [.57, .72], $p < .001$. Standardized coefficients for each individual predictor show that faces were more likely to be judged as human if they were more proportional, alive in the eyes, and familiar; and less memorable, symmetrical, attractive, and smooth-skinned (Table 2).

Which attributes contribute to AI hyperrealism?

Here, we take a novel approach by applying a Brunswikian lens model (Brunswik, 1956; Hall et al.,

Table 1. Experiment 2 Visual-Attribute Rating Conditions

Attribute	N raters for M/F faces	Rating question	Low anchor = 0	High anchor = 100
Age	23/21	How old is this person?	0 years	100 years
Alive in the eyes/ uncanny valley ^a	22/21	When you look at this person's eyes, how alive do they seem?	Definitely not alive	Definitely alive
Attractive	20/22	How attractive is this face?	Not at all	Very attractive
Congruent lighting	21/21	How much would this face stand out in a crowd?	Not at all	Very much
Distinctive/average	22/20	How congruent are the lighting and shadows across this picture?	Not at all	Very congruent
Expressive	21/22	How emotionally expressive is this face?	Not at all	Very expressive
Eye contact	22/22	Is this person making eye contact with you?	No, they are definitely not	Yes, they definitely are
Familiar	21/23	How familiar is this face?	Not at all	Very familiar
Genuinely happy	21/21	How happy is this person genuinely feeling?	Not at all	Very much
Image quality	21/21	How good is the quality of this picture?	Poor	Excellent
Memorable	21/22	How memorable is this face?	Not at all	Very memorable
Proportional/features work as a whole ^b	21/22	How proportional is this face?	Not at all	Very proportional
Smooth-skinned/ perfectness ^c	24/23	How smooth is this person's skin?	Not at all	Very smooth
Symmetrical	22/27	How symmetrical is this face?	Not at all	Very symmetrical

Note: See also the full experimental surveys on the Open Science Framework (osf.io/sz2fe/). ^a“Alive in the eyes” combines the “eyes” and “uncanny valley” themes from Experiment 1’s qualitative framework. ^b“Proportional” combines the “features work as a whole” (are in proportion with one another) and “proportional” themes. ^c“Smooth-skinned” derives from “the skin or wrinkles” and “perfectness” themes.

2019) to reveal how each of the 14 attributes contributed to faces being (mis)judged as human (Fig. 5). Constructing a stimulus-level lens model (using *lavaan*; Rossell, 2012) allowed us to investigate the attributes as simultaneous mediators explaining the correspondence between the AI or human status of faces and how often they were judged as human, thereby distinguishing between cue validity (differences in the visual attributes of human and AI faces) and cue utilization (the extent to which each attribute contributes to faces being judged as human). In this model, face type was the focal predictor (AI = 0 vs. human = 1), the 14 attributes were entered as simultaneous mediators, and the percentage of participants who judged each face as human was the outcome variable. All attributes were allowed to freely correlate with each other (see Supplemental File S12 for the correlation matrix), producing a fully saturated model that perfectly fit the data (comparative fit indices = 1.00, Tucker-Lewis indices = 1.00, root-mean-square errors of approximation = .00).³ Consistent with Experiment 1, the total effect indicated

that AI faces were more likely to be perceived as human than actual human faces, $\beta = -0.41$, $z = -6.26$, $p < .001$.

Critically, in line with our face-space theory prediction that AI faces would be more average than human ones, AI faces were significantly more average (less distinctive), familiar, and attractive, and less memorable than human faces. Overall, AI hyperrealism was explained by larger cumulative effects for the attributes that were utilized in the wrong direction—facial proportions, familiarity, and memorability (in red, Fig. 5; $\beta = -0.67$, 95% CI = $[-.88, -.46]$, $z = -6.10$, $p < .001$)—compared with those utilized in the correct direction—facial attractiveness, symmetry, and congruent lighting/shadows (in green; $\beta = 0.37$, 95% CI = $[.21, .53]$, $z = 4.47$, $p < .001$). Additionally, several valid cues were not utilized by participants—namely, facial averageness/distinctiveness, image quality, and expressivity (in gray). There also remained a residual direct effect of face type (human vs. AI) on perceiving the face as human, over and above the 14 attributes we measured, $\beta = -0.22$, 95% CI = $[-.40, -.03]$, $z = -2.33$, $p = .020$. Therefore,

Table 2. Standardized Coefficients for Each Attribute (Ordered by β Weight) in Our Linear Regression Model Predicting Experiment 1 Stimulus-Level Percentage Judged as Human

Attribute	β	<i>SE</i>	<i>t</i>	<i>p</i>	95% CI
Proportional	0.67	0.11	6.05	< .001***	[0.45, 0.89]
Alive in the eyes	0.37	0.08	4.56	< .001***	[0.21, 0.53]
Expressive	0.23	0.14	1.68	.096†	[-0.04, 0.51]
Familiar	0.20	0.08	2.69	.008**	[0.05, 0.35]
Eye contact	-0.01	0.05	-0.17	.864	[-0.11, 0.09]
Distinctive/average	-0.04	0.08	-0.45	.654	[-0.19, 0.12]
Image quality	-0.08	0.07	-1.10	.274	[-0.22, 0.06]
Congruent lighting	-0.10	0.06	-1.63	.104	[-0.22, 0.02]
Age	-0.14	0.10	-1.42	.156	[-0.33, 0.05]
Memorable	-0.17	0.08	-2.17	.031*	[-0.32, -0.02]
Symmetrical	-0.21	0.09	-2.33	.021*	[-0.39, -0.03]
Attractive	-0.27	0.09	-2.90	.004**	[-0.45, -0.09]
Genuinely happy	-0.28	0.15	-1.81	.071†	[-0.58, 0.02]
Smooth-skinned	-0.61	0.10	-6.07	< .001***	[-0.81, -0.41]

Note: CI = confidence interval; boldface type indicates $p < .05$. * $p < .05$. ** $p < .01$. *** $p < .001$.

† $p < .10$.

although our identified mediators explained the majority of the AI hyperrealism effect, there are aspects still to be uncovered (we return to this in the Discussion).

Can human-perceived attributes be used to accurately classify AI and human faces?

Given that humans are unable to detect current AI faces, society needs tools that can accurately identify AI imposters. Present AI detection algorithms are limited to specific databases (e.g., the popular Google Chrome extension, V7 Fake Profile Detector, works only for StyleGAN faces). Human perception may be useful for improving algorithmic generalizability, as integrating additional parameters into algorithms has proved useful in other domains (J. W. Miller et al., 2022). We therefore provide the first investigation of whether machine learning can leverage human-perceived attributes to accurately classify AI and human faces.

Using 10-fold cross-validation, we constructed a random forest classification model ($mtry = 4$; the square root of the number of predictors, rounded to the nearest whole number) predicting face type (AI vs. human) from the 14 attributes identified in Experiment 2. The model was able to accurately classify face type with 94% accuracy, 95% CI = [91%, 97%], $z = 56.53$, $p < .001$, $\kappa = .88$ (also see Table 3 for the confusion matrix specifying the predicted and actual values during cross-validation). AI faces, at least those generated by StyleGAN2, can therefore be distinguished from human faces

on the basis of human-perceived attributes with extremely high accuracy.

General Discussion

We find that White AI faces are perceived as hyperreal and that observers are overconfident in their ability to detect them. By combining psychological theory with a novel data-driven approach and machine learning, our study significantly advances understanding of why AI hyperrealism occurs. Specifically, we were able to pinpoint perceptual attributes that accurately distinguish AI from human faces and model how people misuse this information, explaining a significant majority of the variance in humans' AI judgments. The identification of these attributes provides a critical foundation in the future for detailed psychophysics work aiming to map AI face-space. Importantly, the present findings are generalizable to the types of images used online, because AI faces are screened for image artifacts as they are selected for real-world use (e.g., when committing fraud; Satter, 2021). Also, artifact screening cannot explain the White specificity of hyperrealism in our reanalysis of Nightingale and Farids's (2022) data, as the same screening criteria were applied across face race.

Our study highlights two separate, and critical, biases. First, generative adversarial networks (GANs) are biased toward the statistical regularities of their most common inputs, which we argue produces AI hyperrealism. Although we demonstrate this point in

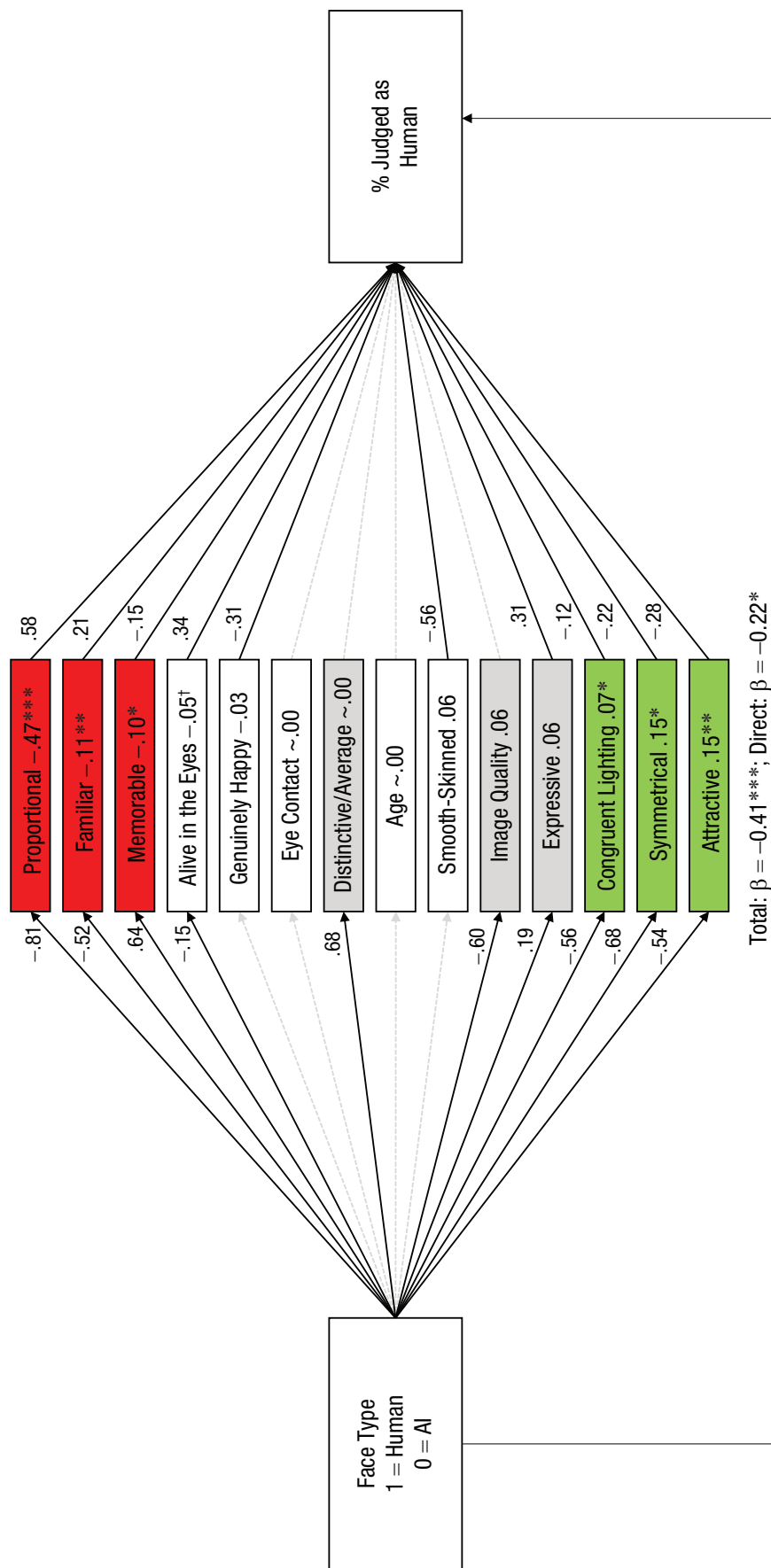


Fig. 5. Lens model testing contributions of each attribute to (mis)judgment of faces as human (ordered by indirect effect size). Red boxes show significant negative indirect effects—attributes that were utilized in the wrong direction to judge AI/human status. Green boxes show significant positive indirect effects (attributes that contributed to accurate AI/human judgments). Gray boxes show attributes that are useful for detecting AI faces but were not utilized by human observers. Dashed lines indicate nonsignificant effects (see Table S13 in the Supplemental Material).
 $^*p < .05$. $^{**}p < .01$. $^{***}p < .001$.

the context of AI faces, the foundational idea may generalize to other important types of AI outputs, including text and artwork from ChatGPT and DALL-E. Here, as this argument predicts, we found AI faces appeared more average than their human counterparts (see Supplemental Fig. S14). Notably, participants failed to utilize facial distinctiveness/averageness for AI detection and inappropriately utilized several other associated cues (facial proportions, memorability, familiarity), producing hyperrealism. Attractiveness was correctly used as a distinguishing feature, confirming Tucciarelli et al.’s (2022) initial finding. The minority of variance left to be explained suggests that other cues, such as those mentioned less often in Experiment 1 (e.g., “ears,” “glasses”), may also play a small but cumulative role.

Second, we found evidence of White racial bias in algorithmic training that produces racial differentials in the presence of AI hyperrealism, with significant implications for the use of AI faces online and in science. Previously, less realistic computer-generated faces have been used as stand-ins for human faces when it was inappropriate to do so (Dawel et al., 2022; E. J. Miller et al., 2023), and there is concern that the same will happen with AI faces, with implications for inequality. We recommend that studies using AI faces should verify that they are perceived as equally natural across races. On a related note, a pressing question is how to address racial differentials in GANs. It is unclear in face-space theory whether there is one face-space or separate spaces for different demographic groups (e.g., Valentine et al., 2016). Future research could fruitfully test these theoretical questions by comparing a GAN trained on equal numbers of faces of each race with GANs trained separately on different demographic groups.

Importantly, and in contrast to standard AI detection algorithms (which are “black boxes”), the present work makes known the perceptual attributes that lead to accurate AI detection in machine learning. Human accuracy may also be improved by training people to utilize attributes appropriately, though this strategy risks exacerbating overconfidence as technologies progress and certain attributes become outdated. Currently, most algorithms produce only single images of each identity, but soon multiple images of AI products are likely to be available (Chan et al., 2023). We likewise drew on a theoretical account of face-space that focuses on variation between single images; when multiple within-identity AI images are commonplace, future work could apply more nuanced face-space theories (e.g., Burton et al., 2016; O’Toole et al., 2018). Regardless, because AI technology is advancing so rapidly (Bond, 2023), training focused on metacognition and education may be more helpful. For example, Szpitalak

Table 3. Confusion Matrix of Correct and Incorrect Machine Classifications

	Actual: AI	Actual: Human
Predicted: AI	93	5
Predicted: Human	7	95

Note: Correct classifications are in boldface.

et al. (2021) found that people who were advised about the unreliability of human memory were more resistant to misinformation than naive individuals. Educating people about the perceived realism of AI faces could likewise reduce risks by making the public appropriately skeptical.

We also found individual differences in the accuracy of AI face detection (Fig. 2c), opening new lines of research. Participants were selected to have normal-range face perception, yet the best performer achieved only 80% accuracy. However, people with exceptional face recognition abilities (super recognizers; Ramon et al., 2019) may possess superior AI detection skills. A further intriguing question is whether individual differences in the utilization of specific attributes can shed light on why certain individuals are more vulnerable to deception by AI faces.

Conclusion

The present study demonstrates a robust AI hyperrealism effect: Remarkably, White AI faces can convincingly pass as more real than human faces—and people do not realize they are being fooled. We believe psychology has a critical role to play in holding AI technologies accountable to the public good. Society has faced many large-scale, seemingly unsolvable challenges that have subsequently become a normal, and manageable, part of life (e.g., automobile safety). We remain hopeful that social and regulatory responses will reduce potential risks as society adjusts to the inevitable presence of AI in our world.

Transparency

Action Editor: Rachael Jack
Editor: Jennifer L. Tackett
Author Contributions
Elizabeth J. Miller: Conceptualization; Data curation; Formal analysis; Methodology; Visualization; Writing – original draft; Writing – review & editing.
Ben A. Steward: Formal analysis; Methodology; Visualization; Writing – original draft; Writing – review & editing.
Zak Witkower: Formal analysis; Methodology; Visualization; Writing – original draft; Writing – review & editing.

Clare A. M. Sutherland: Conceptualization; Funding acquisition; Methodology; Writing – review & editing.

Eva G. Krumhuber: Conceptualization; Funding acquisition; Methodology; Writing – review & editing.

Amy Dawel: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration; Supervision; Visualization; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests


The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.


Funding

This research is supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (Project No. DP220101026), a TRANSFORM Career Development Fellowship to A. Dawel from the Australian National University College of Health and Medicine, and an Experimental Psychology Society Small Grant to C. A. M. Sutherland. The funders had no role in developing or conducting this research.



ORCID iDs

Elizabeth J. Miller  <https://orcid.org/0000-0003-2572-6134>

Ben A. Steward  <https://orcid.org/0000-0002-7517-9215>

Zak Witkower  <https://orcid.org/0000-0002-6767-9834>

Clare A. M. Sutherland  <https://orcid.org/0000-0003-0443-3412>

Eva G. Krumhuber  <https://orcid.org/0000-0003-1894-2517>

Amy Dawel  <https://orcid.org/0000-0001-6668-3121>

Acknowledgment

We thank Sophie J. Nightingale and Hany Farid for providing open access to their stimuli and data.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976231207095>

Notes

1. Nightingale and Farid's (2022) Experiment 1 sample was 73% White. We ran a mixed analysis of variance on the percentage of faces judged as human with face type (White AI, White human) within subjects and participant race (White, non-White) between subjects and found no significant main effect of participant race, $F(1, 313) = 1.23$, $MSE = 267.22$, $p = .268$, or interaction with face type, $F(1, 313) = 1.01$, $MSE = 581.35$, $p = .316$. However, the other-race effect arises from a lack of early-life exposure to other-race faces (McKone et al., 2019; Singh et al., 2022), which may be unevenly distributed across non-White participants. Therefore, we took a conservative approach and recruited only White participants in the current studies.

2. Constructing a binomial regression model instead of a linear one yielded a nearly identical pattern of results (Supplemental File S11).

3. We were interested in whether the attributes differed for human versus AI faces and whether each attribute relates to perceptions of faces being human rather than to how these attributes relate to each other. By allowing all attributes to correlate, we excuse model-fit issues generated by these expected covariances.

References

- Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, 16(3), Article 40. <https://doi.org/10.1167/16.3.40>
- Bond, S. (2023, March 23). It takes a few dollars and 8 minutes to create a deepfake. And that's only the start. *NPR*. <https://www.npr.org/2023/03/23/1165146797/it-takes-a-few-dollars-and-8-minutes-to-create-a-deepfake-and-thats-only-the-sta>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Quantitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.
- Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, 40(1), 202–223. <https://doi.org/10.1111/cogs.12231>
- Chan, E. R., Nagano, K., Chan, M. A., Bergman, A. W., Park, J. J., Levy, A., Aittala, M., De Mello, S., Karras, T., & Wetzstein, G. (2023, April 5). Generative novel view synthesis with 3D-aware diffusion models. *ArXiv.org*. <https://arxiv.org/abs/2304.02602v1>
- Chandaliya, P. K., & Nain, N. (2022). ChildGAN: Face aging and rejuvenation to find missing children. *Pattern Recognition*, 129, Article 108761. <https://doi.org/10.1016/j.patcog.2022.108761>
- Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2022). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, 54(4), 1889–1901. <https://doi.org/10.3758/s13428-021-01705-3>
- Devlin, H. (2023, May 3). AI 'could be as transformative as Industrial Revolution.' *The Guardian*. <https://www.theguardian.com/technology/2023/may/03/ai-could-be-as-transformative-as-industrial-revolution-patrick-vallance>
- Flowe, H. D., Humphries, J. E., Takarangi, M. K., Zelek, K., Karoğlu, N., Gabbert, F., & Hope, L. (2019). An experimental examination of the effects of alcohol consumption and exposure to misleading postevent information on remembering a hypothetical rape scenario. *Applied Cognitive Psychology*, 33(3), 393–413. <https://doi.org/10.1002/acp.3531>
- Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Grope, D. M., Khuvis, S., Herrero, J. L., Irani, M., Mehta, A. D., & Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, 10(1), Article 1. <https://doi.org/10.1038/s41467-019-12623-6>

- Hall, J. A., Horgan, T. G., & Murphy, N. A. (2019). Nonverbal Communication. *Annual Review of Psychology*, 70(1), 271–294. <https://doi.org/10.1146/annurev-psych-010418-103145>
- Hao, K. (2021, June 11). These creepy fake humans herald a new age in AI. *MIT Technology Review*. <https://www.technologyreview.com/2021/06/11/1026135/ai-synthetic-data/>
- Hatmaker, T. (2020, September 22). Chinese propaganda network on Facebook used AI-generated faces. *TechCrunch*. <https://techcrunch.com/2020/09/22/facebook-gans-takes-down-networks-of-fake-accounts-originating-in-china-and-the-philippines/>
- JASP Team. (2023). *JASP* (0.17) [Computer software]. <https://jasp-stats.org/>
- Karras, T., Laine, S., & Aila, T. (2021). A style-based generator architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12), 4217–4228. <https://doi.org/10.1109/TPAMI.2020.2970919>
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of StyleGAN*. https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta- d' , response-specific meta- d' , and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3
- McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., Kidd, E., & Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports*, 9(1), Article 12820. <https://doi.org/10.1038/s41598-019-49202-0>
- McLoughlin, N., Tipper, S. P., & Over, H. (2018). Young children perceive less humanness in outgroup faces. *Developmental Science*, 21(2), Article e12539. <https://doi.org/10.1111/desc.12539>
- Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1), 3–35. <https://doi.org/10.1037/1076-8971.7.1.3>
- Miller, E. J., Foo, Y. Z., Mewton, P., & Dawel, A. (2023). How do people respond to computer-generated versus human faces? A systematic review and meta-analyses. *Computers in Human Behavior Reports*, 10, Article 100283. <https://doi.org/10.1016/j.chbr.2023.100283>
- Miller, J. W., O'Neill, C., Constantinou, N. C., & Azencot, O. (2022, December 23). Eigenvalue initialisation and regularisation for Koopman autoencoders. *ArXiv.org*. <https://arxiv.org/abs/2212.12086v2>
- Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences, USA*, 119(8), Article e2120481119. <https://doi.org/10.1073/pnas.2120481119>
- O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, 22(9), 794–809. <https://doi.org/10.1016/j.tics.2018.06.006>
- Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., Albonico, A., Malaspina, M., Daini, R., Irons, J., Al-Janabi, S., Taylor, L. C., Rivolta, D., & McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology*, 70(2), 218–233. <https://doi.org/10.1080/17470218.2016.1161058>
- Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences, USA*, 119(17), Article e2115228119. <https://doi.org/10.1073/pnas.2115228119>
- Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, 110(3), 461–479. <https://doi.org/10.1111/bjop.12368>
- R Core Team. (2021). *R: A language and environment for statistical computing* (4.2.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, 57(1), 199–226. <https://doi.org/10.1146/annurev.psych.57.102904.190208>
- Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, 46(18), 2977–2987. <https://doi.org/10.1016/j.visres.2006.03.002>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Satter, R. (2021, April 20). Experts: Spy used AI-generated face to connect with targets. *AP News*. <https://apnews.com/article/ap-top-news-artificial-intelligence-social-platforms-think-tanks-politics-bc2f19097a4c4fffaa00de6770b8a60d>
- Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021). A study of the human perception of synthetic faces. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1–8). <https://doi.org/10.1109/FG52635.2021.9667066>
- Singh, B., Mellinger, C., Earls, H. A., Tran, J., Bardsley, B., & Correll, J. (2022). Does cross-race contact improve cross-race face perception? A meta-analysis of the cross-race deficit and contact. *Personality and Social Psychology Bulletin*, 48(6), 865–887. <https://doi.org/10.1177/01461672211024463>

- Szpitalak, M., Woltmann, A., Polczyk, R., & Kękuś, M. (2021). Memory training as a method for reducing the misinformation effect. *Current Psychology*, 40(11), 5410–5419. <https://doi.org/10.1007/s12144-019-00490-9>
- Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (2022). On the realness of people who do not exist: The social processing of artificial faces. *IScience*, 25(12), Article 105441. <https://doi.org/10.1016/j.isci.2022.105441>
- Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 43(2), 161–204. <https://doi.org/10.1080/14640749108400966>
- Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, 69(10), 1996–2019. <https://doi.org/10.1080/17470218.2014.990392>
- Vokey, J. R., & Read, J. D. (1992). Familiarity, memorability, and the effect of typicality on the recognition of faces. *Memory & Cognition*, 20(3), 291–302. <https://doi.org/10.3758/BF03199666>
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18(1), 10–65. <https://doi.org/10.1177/1529100616686966>
- Xie, H. (2023). The promising future of cognitive science and artificial intelligence. *Nature Reviews Psychology*, 2(4), Article 4. <https://doi.org/10.1038/s44159-023-00170-3>