Supplemental Materials for:

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

Elizabeth J. Miller[1], Ben A. Steward[1], Zak Witkower[2], Clare A. M. Sutherland[3,4], Eva G. Krumhuber[5], Amy Dawel[1]

Published in *Psychological Science*, a journal of the Association for Psychological Science.

[1]School of Medicine and Psychology, Australian National University, Canberra, Australia.

[2]Department of Psychology, University of Toronto, Toronto, Canada.

[3]School of Psychology, King's College, University of Aberdeen, Aberdeen, Scotland.

[4]School of Psychological Science, University of Western Australia, Crawley, WA, Australia.

[5]Department of Experimental Psychology, University College London, UK.

- Elizabeth J. Miller liz.miller@anu.edu.au ORCiD: 0000-0003-2572-6134
- Ben A. Steward ben.steward@anu.edu.au ORCiD: 0000-0002-7517-9215
- Zak Witkower zakwitkower@gmail.com ORCiD: 0000-0002-6767-9834
- Clare A. M. Sutherland clare.sutherland@abdn.ac.uk ORCiD: 0000-0003-0443-3412
- Eva G. Krumhuber e.krumhuber@ucl.ac.uk ORCiD: 0000-0003-1894-2517
- Amy Dawel amy.dawel@anu.edu.au ORCiD: 0000-0001-6668-3121

Author Note

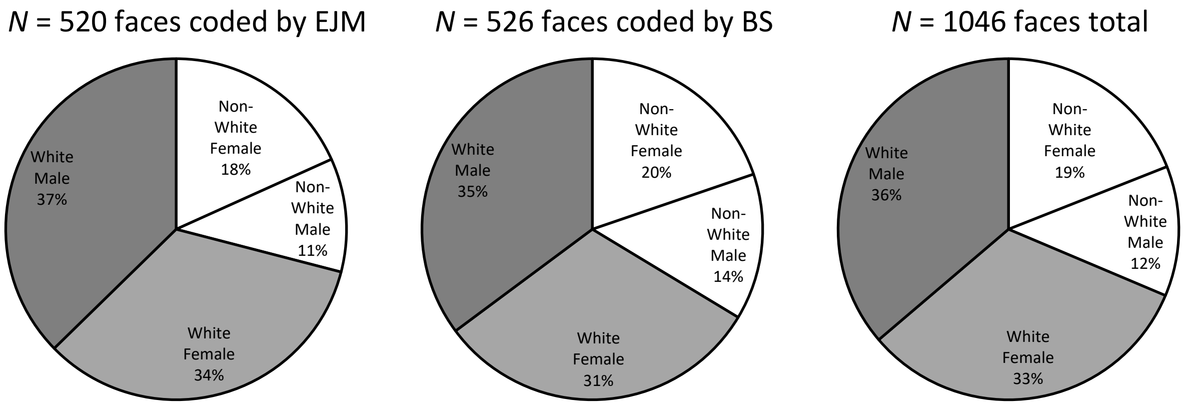
This research is supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP220101026), a TRANSFORM Career Development Fellowship to AD from The Australian National University (ANU) College of Health and Medicine, and an Experimental Psychology Society Small Grant to CS. The funders had no role in developing or conducting this research. We have no conflicts of interest to disclose. *Correspondence concerning this article should be addressed to Amy Dawel, School of Medicine and Psychology (building 39), The Australian National University, Canberra, ACT 2600, Australia. Email: amy.dawel@anu.edu.au

Table of Contents

Supplement S1 — Demographics of the Faces in the Flickr-Faces-HQ Dataset (Karras et al., 2021) used to Train StyleGAN2

Figure S1

Demographics of the Faces in the Flickr-Faces-HQ Dataset (Karras et al., 2021) used to Train StyleGAN2



N = 520 faces coded by EJM

N = 526 faces coded by BS

N = 1046 faces total

Note. Demographics are for 1,046 faces selected at random from the Flickr-Faces-HQ Dataset (Karras et al., 2021). EJM and BS each independently coded ~half of the faces to verify these demographic differentials were reliable at this sample size. Seven additional images were excluded from coding because they showed multiple faces or were too blurry to identify the face race.

Supplement S2 — Participant Data Exclusions

Table S2

Participant Data Exclusions for Experiments 1 and 2

Reason	E1	E2
Did not mean inclusion criterion: not White	2	9
Did not mean inclusion criterion: lived > 2 yrs outside US pre-18	3	5
Did not mean inclusion criterion: condition affecting face tasks	36	96
Incomplete[a]	16	75
Completed study on phone	6	36
>1 error on catch trials	18	37
Failed final attention check	0	54[b]
Aware that faces may be AI (in debrief Q re: anything unusual)	N/A	44
Total exclusions *N*	81	356
Retained sample *N*	124	610

Note. [a]Incomplete includes participants who entered the study but did not start the face task. [b]The higher rate of failed final attention checks for Experiment 2 reflects that this check was more difficult than for Experiment 1, requiring participants to select the face task they had completed from a list of 14 rather than 4 options. Many errors reflect confusion amongst closely related tasks (e.g., selecting the expressivity option when the participant had completed the “How happy is this person genuinely feeling?” task).

Supplement S3 — Power and Sensitivity Analyses

We used G*Power version 3.1 (Faul et al., 2009) for all power and sensitivity analyses, with power set to 0.95 and α set to .05.

Experiment 1. An a priori power analysis was conducted for detecting correlations between judgment accuracy and confidence ratings, because detecting correlations required a larger sample size than detecting accuracy differences between AI and human faces. Using the exact correlation bivariate normal model one-tailed option (because we had a one-way a priori hypothesis that higher accuracy would be associated with higher confidence) in G*Power showed that detecting correlations ≥ .3 (a medium effect size) required a sample size of 115 participants, which we exceeded with our sample of 124.

Experiment 2. An a priori sensitivity analysis was conducted for predicting the percentage of participants who had categorized each face as human using mean physical attribute ratings for each stimulus. The sample size was the number of faces and was therefore predetermined as 200. Using the linear multiple regression fixed model R^2 increase option in G*Power showed that our study was powered at 0.95 to detect a small to medium effect size of f^2 = 0.066 when testing 1 of a total 14 predictors. Separately, the sample size for each rating type was determined by the need to obtain reliable mean ratings for each stimulus. DeBruine & Jones (2022) estimate that 15 raters are required to achieve Cronbach's α > .8 for attractiveness. We took a conservative approach and recruited a minimum of 20 raters per stimulus for each of the 14 attributes of interest. Table S3 shows Cronbach's α for ratings of each attribute in Experiment 2 averaged .95 across the attributes and face types and was > .84 in all instances, indicative of excellent reliability.

Table S3

Cronbach's αs for the 14 Attributes Rated in Experiment 2 for Each Face Type by Face Sex

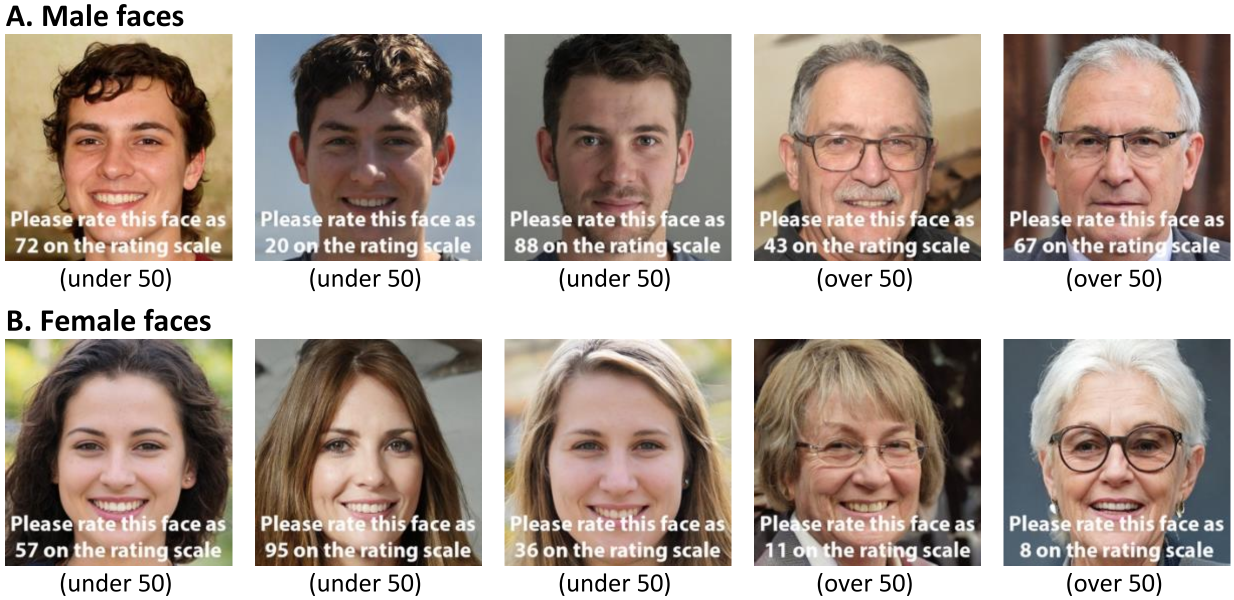
	AI				Human			
	Male		Female		Male		Female	
Attribute	alpha	95% CI	alpha	95% CI	alpha	95% CI	alpha	95% CI
Age	0.90	[0.86, 0.94]	0.94	[0.91, 0.96]	0.86	[0.80, 0.91]	0.92	[0.88, 0.95]
Alive in the eyes/uncanny valley	0.96	[0.94, 0.97]	0.98	[0.96, 0.98]	0.95	[0.92, 0.97]	0.97	[0.95, 0.98]
Attractive	0.98	[0.97, 0.99]	0.96	[0.94, 0.97]	0.98	[0.97, 0.99]	0.97	[0.95, 0.98]
Distinctive/average	0.97	[0.96, 0.98]	0.98	[0.98, 0.99]	0.97	[0.95, 0.98]	0.95	[0.93, 0.97]
Expressive	0.85	[0.77, 0.90]	0.94	[0.91, 0.96]	0.85	[0.78, 0.91]	0.90	[0.85, 0.94]
Eye contact	0.95	[0.92, 0.97]	0.95	[0.93, 0.97]	0.97	[0.95, 0.98]	0.96	[0.94, 0.97]
Familiar	0.99	[0.98, 0.99]	0.99	[0.99, 0.99]	0.98	[0.97, 0.99]	0.96	[0.94, 0.98]
Genuinely happy	0.97	[0.95, 0.98]	0.96	[0.94, 0.97]	0.96	[0.94, 0.98]	0.95	[0.92, 0.97]
Image quality	0.98	[0.97, 0.99]	0.98	[0.96, 0.98]	0.97	[0.95, 0.98]	0.97	[0.96, 0.98]
Congruent lighting	0.93	[0.89, 0.96]	0.96	[0.95, 0.98]	0.92	[0.88, 0.95]	0.93	[0.89, 0.96]
Memorable	0.97	[0.96, 0.98]	0.97	[0.95, 0.98]	0.98	[0.97, 0.99]	0.91	[0.86, 0.94]
Proportional/features work as a whole	0.96	[0.94, 0.98]	0.97	[0.95, 0.98]	0.97	[0.95, 0.98]	0.97	[0.96, 0.98]
Smooth skinned	0.97	[0.96, 0.98]	0.96	[0.94, 0.97]	0.97	[0.96, 0.98]	0.96	[0.94, 0.97]
Symmetrical	0.96	[0.95, 0.98]	0.96	[0.95, 0.98]	0.95	[0.93, 0.97]	0.96	[0.94, 0.98]

Supplement S4 — Catch Trial Stimuli

Figure S4 shows the stimuli used for attention catch trials. These stimuli were generated using thispersondoesnotexist.com which uses the StyleGAN2 algorithm (Karras et al., 2020, 2021), to be obviously under or over 50 years old for the catch trials in Experiment 1. The same stimuli were used in Experiment 2 overlaid with the requested numerical rating for each catch trial, as shown here.

Figure S4

Catch Trial Stimuli


A. Male faces
Please rate this face as 72 on the rating scale
(under 50)
Please rate this face as 20 on the rating scale
(under 50)
Please rate this face as 88 on the rating scale
(under 50)
Please rate this face as 43 on the rating scale
(over 50)
Please rate this face as 67 on the rating scale
(over 50)
B. Female faces
Please rate this face as 57 on the rating scale
(under 50)
Please rate this face as 95 on the rating scale
(under 50)
Please rate this face as 36 on the rating scale
(under 50)
Please rate this face as 11 on the rating scale
(over 50)
Please rate this face as 8 on the rating scale
(over 50)

Note. The white text appeared on the faces in Experiment 2 but not Experiment 1, which instead asked participants to judge if these faces were over or under 50 years old.

Supplement S5 — Participant Means and Comparisons

Table S5

Wilcoxon Signed-rank Tests[a] Comparing M Percent Human Judgments and M Ratings for AI and Human Face Participant Means

Variable	*N*	*M* AI	*M* human	*z*	*p*	*Rank-biserial correlation*
Human sig. > AI						
E2: *M* expressive	43	48.3	56.1	5.65	**< .001**	0.989
E2: *M* genuinely happy	42	51.0	55.2	4.80	**< .001**	0.849
E2: *M* eye contact	44	65.5	71.7	4.89	**< .001**	0.846
E2: *M* distinctive/average	42	39.9	52.6	4.41	**< .001**	0.782
E2: *M* memorable	43	43.9	55.3	4.29	**< .001**	0.751
AI sig. > human						
E2: *M* symmetrical	49	55.8	51.4	-6.09	**< .001**	-1.000
E2: *M* proportional	43	71.1	50.2	-5.71	**< .001**	-1.000
E2: *M* attractive	42	56.7	41.6	-5.62	**< .001**	-0.996
E2: *M* congruent lighting	42	62.8	47.3	-5.55	**< .001**	-0.982
E2: *M* image quality	42	72.8	62.0	-5.39	**< .001**	-0.949
E2: *M* familiar	44	31.6	21.9	-4.88	**< .001**	-0.844
N&F E1: % human	315	69.5	52.2	-10.84	**< .001**	-0.736
E2: *M* smooth skinned	47	64.8	60.6	-4.13	**< .001**	-0.691
E2: *M* age	44	34.6	33.5	-3.29	**< .001**	-0.570
E1: % human	124	65.9	51.1	-4.40	**< .001**	-0.459
No sig. diff between AI & human[a]						
E1: *M* confidence	124	64.8	64.6	0.33	.742	0.034
E2: *M* alive in the eyes	43	67.8	64.7	-1.81	.072	-0.320

Note. N&F E1 = Nightingale and Farid (2022) Experiment 1. E1 = current Experiment 1. E2 = current experiment 2. [a]Results are for Wilcoxon signed-rank tests, used because many variables were non-normally distributed and/or the AI and human variable distributions violated the assumption of equal variances. [b]α is Bonferroni corrected for the number of comparisons within each experiment. For example, E2 = 14 comparisons and therefore $\alpha = .05/14 = .004$. Bold text indicates $p <$ Bonferroni-corrected α.

Supplement S6 — Stimulus Means and Comparisons

Table S6

Welch's t-tests[a] Comparing M Percent Human Judgments and M Ratings for AI and Human Face Stimulus Means (N = 200 stimuli)

Variable	*M* AI	*M* human	*t*	df	*p*	*d*
Human sig. > AI						
E2: *M* distinctive/average	39.9	52.6	13.09	190	**< .001**	1.85
E2: *M* memorable	44.0	55.2	11.72	184	**< .001**	1.66
E2: *M* expressive	48.3	56.1	2.77	196	**0.006**	0.39
AI sig. > human						
E2: *M* proportional	71.1	50.2	-19.70	192	**< .001**	-2.79
E2: *M* symmetrical	66.7	51.4	-13.11	196	**< .001**	-1.85
E2: *M* image quality	72.8	62.0	-10.63	193	**< .001**	-1.50
E2: *M* congruent lighting	62.8	47.2	-9.41	193	**< .001**	-1.33
E2: *M* attractive	56.5	41.6	-8.92	194	**< .001**	-1.26
N&F E1: % human	70%	52%	-8.53	160	**< .001**	-1.21
E2: *M* familiar	31.4	21.9	-8.45	197	**< .001**	-1.20
E1: % human	66%	51%	-6.23	189	**< .001**	-0.88
No sig. diff between AI & human[a]						
E2: *M* eye contact	65.5	71.7	1.72	192	0.087	0.24
E2: *M* genuinely happy	51.0	55.2	1.55	196	0.124	0.22
E1: *M* confidence	64.8	64.6	-0.41	195	0.684	-0.06
E2: *M* age	34.5	33.5	-0.50	196	0.616	-0.07
E2: *M* smooth skinned	65.0	60.6	-1.60	197	0.112	-0.23
E2: *M* alive in the eyes	67.8	64.8	-2.10	196	0.037	-0.30

Note. N&F E1 = Nightingale and Farid (2022) Experiment 1. E1 = current Experiment 1. E2 = current experiment 2. [a]Results are for Welch's *t*-tests, used because many variables were non-normally distributed and/or the AI and human variable distributions violated the assumption of equal variances. [b]α is Bonferroni corrected for the number of comparisons within each experiment. For example, E2 = 14 comparisons and therefore α = .05/14 = .004. Bold text indicates *p* < Bonferroni-corrected α.

Supplement S7 — Experiment 1 Meta-*d′* Analysis

Meta-*d′* is a model-based approach used to describe how well a person's confidence ratings distinguish between their correct and incorrect judgments (known as metacognitive sensitivity), while also accounting for how well that person is performing on the judgment task (Maniscalco & Lau, 2012, 2014). Because meta-*d′* accounts for task accuracy, two people who perform differently on a judgment task can have the same (high) meta-*d′* value, indicating that, for example, person A—a high performer—knows that they are performing well (and therefore gives high confidence ratings) whereas person B—a poor performer—knows they are performing poorly (and therefore gives low confidence ratings).

As both meta-*d′* and *d*' are reported in the same units, it is possible to interpret meta-*d′* as we do *d′*. For example, we can interpret meta-*d′* as how well confidence ratings distinguish between correct and incorrect judgments. Larger positive meta-d' values indicate a person has better insight into their own performance, whereas meta-d' values approaching zero indicate confidence ratings do not distinguish between correct and incorrect judgments and therefore that a person has little or no insight. Conversely, meta-d' values that are negative indicate that confidence ratings distinguish between correct and incorrect judgments, but in the wrong direction. This situation occurs when a person reports relatively high confidence for incorrect judgments and relatively low confidence for correct judgments, and reflects particularly poor insight—as we find for the majority of participants in Experiment 1.

Calculation of Meta-*d′*

Meta-*d′* was calculated in Matlab using code provided by Maniscalco & Lau (2012, 2014; code available at: http://www.columbia.edu/~bsm2105/type2sdt/). Because this calculation requires confidence ratings to be interval-level data, we recoded each participant's confidence ratings from 0 to 100 into ten new confidence variables which were counts of the total number of ratings they gave within each of five bins (0-20, 21-40, 41-60, 61-80, 81-100) for each of the two response categories (AI or human) across all the trials they completed. This was done twice for each participant, once for Human face trials and again for AI face trials. For example, if a participant was presented with an AI face and responded that the face was human and gave a confidence rating of 90, this was counted as 1 for the last bin of the AI trials. Because this binning sometimes produces a zero value, a small correction was made to the data, as recommended by Maniscalco & Lau (2012, 2014), to allow meta-*d′* to be correctly estimated. The correction, calculated as 1 divided by the number of confidence variables (i.e., 1/10 = 0.1), was added to each confidence variable for every participant. These data were then used to calculate meta-*d′*. Outliers were assessed for *meta-d′* using the outlier labelling rule (set at an interquartile range of 2.2; Hoaglin & Iglewicz, 1987). Removing the data of six participants that were identified as outliers marginally reduced the strength of the correlation between meta-*d′* and *d′* from $r = .479$, $p < .001$ to $r = .453$, $p < .001$, but did not otherwise change the pattern of findings and we therefore retained them in the main analyses.

Supplement S8 — Experiment 1 Analyses by Face Sex

Overall, analyses by face sex found that both male and female White AI faces were misjudged as human more often than real human faces and more often than chance, and that this effect tended to be stronger for male than female faces (Figure S9). A mixed ANOVA on the percentage of faces judged as human by each participant, with face type (AI, human) within subjects and face sex (male, female) between subjects, found a significant main effect for face type, *F*(122) = 52.80, *MSE* = 253.81, *p* < .001, η^2_p = 0.30, reproducing the main text finding that AI faces were judged as human more often than real human faces (M_{AI} = 65.9% vs M_{human} = 51.1%). While the main effect of face sex was not significant, *F*(122) = 2.22, *MSE* = 653.13, *p* = .139, η^2_p = 0.02, there was a significant face type by face sex interaction, *F*(122) = 7.10, *MSE* = 253.81, *p* = .009, η^2_p = 0.06. The Bonferroni-corrected posthoc pairwise comparisons in Table S9a show this interaction reflects that AI male faces were judged as human more often than AI female faces, $M_{\text{AI-male}}$ = 70.9% vs $M_{\text{AI-female}}$ = 60.7%, *t* = 2.68, *p* = .048, *d* = 0.48, 95% CI [-0.01, 0.97], as was also found in Nightingale and Farid (2022). However, importantly, the Table S9a posthoc comparisons and one-sample *t*-test results in Table S9b show that both male and female White AI faces were misjudged as human significantly more often than human faces and significantly more often than chance, respectively. *d'* was also significantly negative for male and female faces separately: male faces: *d'* = -0.65 (vs. 0 = no sensitivity), *t*(62) = 8.25, *p* < .001, *d* = 1.04, 95% CI [0.73, 1.34]; female faces: *d'* = -0.33, *t*(60) = 3.29, *p* = .002, *d* = 0.42, 95% CI [0.16, 0.68].

Figure S8

Reanalysis of Data from Experiment 1 of Nightingale and Farid (2022) and Results for Current Experiment 1 by Face Sex

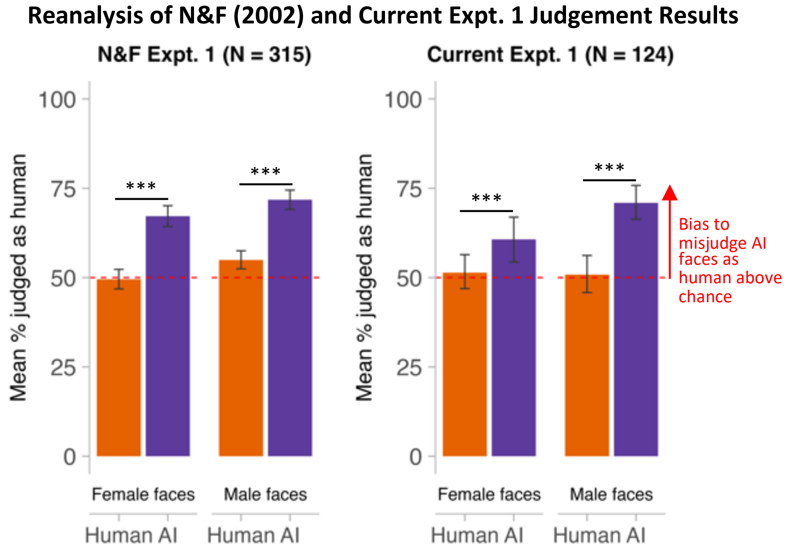


Table S8a

Posthoc Pairwise Comparisons for the Face Type by Face Sex Interaction

Comparison		M_{diff}	95% CI	*SE*	*t*	p_{bonf}	*d*	95% CI
Human female	Human male	0.6%	[-9.6, 10.7]	0.14	0.14	1.000	0.03	[-0.45, 0.51]
Human female	AI female	-9.3%	[-17.1, -1.6]	-3.23	-3.23	**0.010**	-0.44	[-0.81, -0.07]
Human female	AI male	-19.5%	[-29.7, -9.4]	-5.11	-5.11	**< .001**	-0.92	[-1.42, -0.41]
Human male	AI female	-9.9%	[-20.1, 0.3]	-2.58	-2.58	0.064	-0.46	[-0.95, 0.02]
Human male	AI male	-20.1%	[-27.7, -12.5]	-7.08	-7.08	**< .001**	-0.94	[-1.33, -0.55]
AI female	AI male	-10.2%	[-20.4, -0.0]	-2.68	-2.68	**0.048**	-0.48	[-0.97, 0.01]

Note. *p*-value and confidence intervals adjusted for comparing a family of 6 estimates using the Bonferroni method. Bold text indicates *p* < .05.

Table S8b

T-tests Comparing Participants' Percentage of Faces Judged as Human to Chance (= 50%) for Male and Female AI and Human Faces Separately

Face stimulus type	*M*	*t*	df	*p*	*d*	95% CI
AI male	70.9%	8.36	62	**< .001**	1.05	[0.74, 1.36]
AI female	60.7%	3.45	60	**0.001**	0.44	[0.18, 0.70]
Human male	50.8%	0.32	62	0.748	0.04	[-0.21, 0.29]
Human female	51.4%	0.52	60	0.603	0.07	[-0.19, 0.32]

Note. Bold text indicates *p* < .05.

Supplement S9 — Experiment 1 Qualitative Analyses and Results

Most participants were asked what information they used to decide whether faces were AI or human separately, although the first 20 participants answered a single question combining the two. We used χ2 analyses to test if there were significant differences between the expected and observed frequencies in responses coded in each theme for the 104 participants who answered the separate AI and human questions. Adjusted residual values were calculated in SPSS and squared to obtain χ2 values. *p*-values were calculated using the CHISQ.DIST.RT function in Excel and Bonferroni corrected for multiple comparisons by multiplying them by the number of parent nodes in the analysis (i.e., 21). The only significant difference that survived Bonferroni correction in analyses was that participants reported using the skin or wrinkles more for human than AI faces, χ2 = 9.39, *p* = .046. Overall, the results reflect that participants reported using similar information to judge whether faces were AI or human, thereby justifying the integrated thematic framework in Figure 4.

Table S9

Qualitative Coding Framework

Name of theme *(with sub-themes in italics)*	Example quote	Total *N* codes	% human question	% AI question	χ^2	*p*
All seemed human	I did not really think any were computer-generated	6	40.0	60.0	0.12	0.727
Backgrounds	I tried to see if the backgrounds were blurry…	22	33.3	66.7	1.83	0.176
Emotion/expression	…Or the expression that they had, and whether it felt "real"	25	60.9	39.1	1.66	0.198
Facial proportions	If face appeared too wide or elongated…	9	77.8	22.2	3.31	0.069
Features work as a whole	…If part of the face didn't quite go with the rest…	22	36.4	63.6	1.20	0.274
Guessing	…I mostly just guessed	6	33.3	66.7	0.51	0.476
Image quality		59	30.8	69.2	6.70	0.203[a]
Appeared edited	Or whether it looked photoshopped [edited for spelling]	6	50.0	50.0	-	-
Clarity and blur	…If the quality of the photo was low and looked kind of blurry	20	63.2	36.8	-	-
Coloring	I would see slight changes in color…	7	57.1	42.9	-	-
Rendering artefacts	I looked for anything that may be a low-level visual artifact of a computer generation process (small distortions here and there)	26	80.0	20.0	-	-
Instinct or gut feeling	I just mostly went off gut instinct	20	61.1	38.9	1.33	0.249
Jewelry	I also tried to look at things like any jewellery they were wearing [edited for spelling]	6	40.0	60.0	0.12	0.727

Level of detail	Anything that…didn't have enough detail seemed computer-generated	7	71.4	28.6	1.59	0.207
Lighting or shadows	…Were there any weird shadows or were there shadows missing…	29	65.5	34.5	3.88	~1.000[a]
Other		10	50.0	50.0	0.02	0.886
All seemed CG	…they were all AI generated. I'm pretty sure that's correct now - all or at least most were	1	0	100	-	-
Attractiveness	…Their level of attractiveness…	1	0	100	-	-
Child faces	I also assumed most of the pictures of children were computer-generated since I doubt they would be used in the study if they were real humans.	2	100	0	-	-
Features computers struggle to create	If a picture showed teeth I usually put down real human because I find it hard for a computer to replicate disparities in people's teeth.	2	50.0	50.0	-	-
Lack of character	…The lack of character	1	100	0	-	-
Own judgment	Just used my judgment of what looked real and what look artificial	1	0	100	-	-
Seemed dead	If they seemed "dead"… [edited for punctuation]	1	100	0	-	-
Uniqueness of the face	Uniqueness of the face…	1	0	100	-	-
Other physical characteristics		33	53.1	46.9	0.39	0.531
Clothing	I tried to see if any visible clothing looked weird…	5	60.0	40.0	-	-
Glasses	Glasses seemed like a big giveaway, so any time I saw them for the most part I assumed it was a real person	3	33.3	66.7	-	-
Hair or facial hair	…tried to see if I could see individual strands of hair	15	35.7	64.3	-	-
Makeup	Do I think a computer would generate bad makeup?!	2	50.0	50.0	-	-
Shape	Generally looking like a human and looking at the shape of the face…	8	62.5	37.5	-	-
Perfectness or imperfectness		33	35.5	64.5	3.71	0.054
Imperfections signaling humanness	…it looked more real if there were imperfections…	19	5.3	94.7	-	-
Too perfect to be human	I looked for faces that were too perfect	14	83.3	16.7	-	-
Pose or direction of eyes or face	If the eyes were looking off at a strange direction…if the face was aimed in an unnatural direction I figured it was a computer generated image	11	33.3	66.7	0.76	0.382
Real or natural vs artificial		36	36.4	63.6	1.84	0.175
Artificial	Pictures that looked distorted or like a person that was not real…	20	94.7	5.3	-	-
Real or natural	I was looking for faces that seemed natural	16	21.4	78.6	-	-
Repeating features across images	The same set of teeth was used in multiple people, which almost never happens in the natural world	4	50.0	50.0	0.01	0.928
Skin or wrinkles	If the skin looked extremely smooth…	55	67.3	32.7	9.39	**0.046**[a]
Specific facial features		88	43.8	56.3	0.61	0.434
Chin or jaw	Mostly the chin, and the shadow of the chin	3	33.3	66.7	-	-
Ears	Most of the ones I knew were computer-generated had strange ears…	9	77.8	22.2	-	-
Eyes	And no empty gaze in eyes	43	50.0	50.0	-	-
Mouth	There was something about the area around the mouth that felt inhuman…	8	62.5	37.5	-	-
Neck	…if their neck skin looked normal on some of them	3	33.3	66.7	-	-
Noses	I tried to look at noses…	2	0	100.0	-	-
Teeth	If they had super bumpy teeth I put computer-generated	20	63.2	36.8	-	-
Symmetry	Are they too symmetrical? Are they weirdly not symmetrical? [edited for spelling]	41	42.5	57.5	0.48	0.487
Uncanny valley	Something just felt off	24	20.8	79.2	7.32	0.143[a]

Note. Total *N* codes includes codes from all 124 participants, including the 20 participants who answered the single question combining human and AI faces. [a]*p* values < .05 have been Bonferroni corrected for 21 comparisons. All other *p* values are uncorrected (i.e., because corrections often set them to > 1). Bold text indicates *p* < .05.

Supplement S10 — Detailed Rationale for Experiment 2 Visual Attributes

Our predictions and rationale for selecting each of the facial attributes rated in Experiment 2 is detailed below. Our selections were informed by face space theory (Valentine, 1991; Valentine et al., 2016), prior empirical evidence, and the open-ended responses from Experiment 1, targeting the attributes that participants mentioned most often which could be sensibly measured by participant ratings.

Attributes Selected Based on Face Space Theory

(Low) distinctive/(high) average. Our hypothesis that AI faces would cluster around the center of face space predicts that AI faces should be rated as less distinctive/more average than human faces overall, and that the range of distinctiveness ratings should be restricted and towards the lower end of the scale for AI compared to human faces. Indeed, Experiment 2 found AI faces were rated as significantly less distinctive, *M* = 39.9, *SD* = 6.1, range = 27.0 to 61.6, than human faces, *M* = 52.6, *SD* = 7.6, range = 36.8 to 73.8, *t* = 4.41, *p* < .001, *d* = 0.78 (also see Figure S14).

Familiar. We predicted that AI faces would be perceived as more familiar than human ones, as more average faces are perceived as more familiar (Vokey & Read, 1992). We also speculated more familiar faces might be perceived as more human-looking.

Memorable. We predicted that AI faces would be perceived as less memorable than human ones, as more average faces are less memorable (Vokey & Read, 1992).

Attractive. We predicted that AI faces would be perceived as more attractive than human ones, as more average faces are perceived as more attractive (Rhodes, 2006; but cf.

Sofer et al., 2015). We also predicted that more attractive faces would be judged as human less often, following findings from Tucciarelli et al. (2022).

Attributes Selected Based on Experiment 1 Qualitative Responses

Proportional. We predicted that more proportional faces would be judged as human more often because facial proportions can influence perceived humanness (Deska et al., 2018) and 5.6% of Experiment 1 codes referred to features working as a whole (4.0%; e.g., *“if part of a face didn’t quite go with the rest”*) or using facial proportions specifically (1.6%; e.g., *“If face appeared to wide or elongated”*).

Symmetrical. We predicted that more symmetrical faces would be judged as human more often because symmetry is associated with other attributes that may influence perceived humanness—for example, more symmetrical faces are perceived as healthier and more attractive (Rhodes et al., 1998)—and 7.5% of Experiment 1 codes referred to using symmetry.

Alive in the eyes. We predicted that faces that appeared more alive in the eyes would be judged as human more often because removing the corneal reflection from the eyes of a face reduces perceived humanness (Vaitonytė et al., 2021) and 7.9% of Experiment 1 codes referred to using the eyes. We also intended for this attribute to target the uncanny valley feeling of “deadness” captured by 4.4% of Experiment 1 codes.

Eye contact. We predicted that faces making eye contact would be judged as human more often because direct gazing faces are perceived as more human (Khalid et al., 2016) and 7.9% of Experiment 1 codes referred to using the eyes.

Emotional expressivity. We predicted that more expressive faces would be judged as human more often because expressive faces tend to be perceived as more human than non-expressive ones (Bowling & Banissy, 2017; Saito et al., 2022) and 4.6% of Experiment 1 codes referred to using emotional information. While Tucciarelli et al. (2022) found expressivity was not related to judging faces as AI versus human in their study, their stimulus matching procedure included matching for smiles.

Genuinely happy. Following our prediction for emotional expressivity, we predicted that more genuinely happy faces would be judged as human more often. We included ratings of genuine happiness in addition to emotional expressivity because the finding that expressive faces tend to be perceived as more human than non-expressive ones is strongest for happiness (Bowling & Banissy, 2017; Saito et al., 2022) and our visual inspection of the face stimuli found the most common expression was smiling.

Smooth skinned. We predicted that more smooth skinned faces would be judged as human less often because image manipulations that make skin appear more smooth cause faces to be perceived as less human (Vaitonytė et al., 2021, 2022) and 10.1% of Experiment 1 codes referred to using skin texture or wrinkles (e.g., *"If the skin looked extremely smooth", "if the wrinkles looked real"*) and 5.1% of Experiment 1 codes referred to images being too perfect or lacking imperfections (e.g., *"If they looked "too perfect""*).

Congruent lighting and shadows. We predicted that face images with more congruent lighting and shadows would be judged as human more often because 5.3% of Experiment 1 codes referred to using this image attribute in this way (e.g., *“Were [sic] there any weird shadows or were there shadows missing”*).

Image quality. We predicted that face images with more image quality problems would be judged as human less often because when these cues are present people can use them to accurately identify AI images (Tucciarelli et al., 2022) and 10.8% of Experiment 1 codes referred to using image-related properties. Although Nightingale and Farid (2022) screened their images for obvious rendering artefacts, it is possible more subtle cues were overlooked. This attribute also targeted potential background artefacts.

Attributes Selected for Control Purposes

Age. Age was included because we wanted to ensure that any effects involving two related attributes of theoretical and empirical interest—attractiveness and smooth skin—were not better accounted for by age.

Supplement S11 — Binomial Regression Model Results

In parallel with our linear regression, we constructed a binomial regression predicting the percentage of participants who judged a stimulus as human from the 14 stimulus-level attribute means. Overall, the binomial model accounted for 64% of the total variance in how often faces were judged as human, *pseudo* $R^2 = .64$, $p < .001$. Table S11 shows that faces were significantly more likely to be judged as human if they were more proportional, alive in the eyes, and familiar, and less memorable, symmetrical, attractive, smooth skinned, and genuinely happy. Note, this effect for genuinely happy was only marginally significant in the linear regression model reported in the main text. No other differences in statistical significance emerged between the binomial and linear models.

Table S11

Standardized Coefficients for Each Attribute (Ordered by β *Weight) in our Binomial Regression Model Predicting Experiment 1 Stimulus-level Percentage Judged as Human*

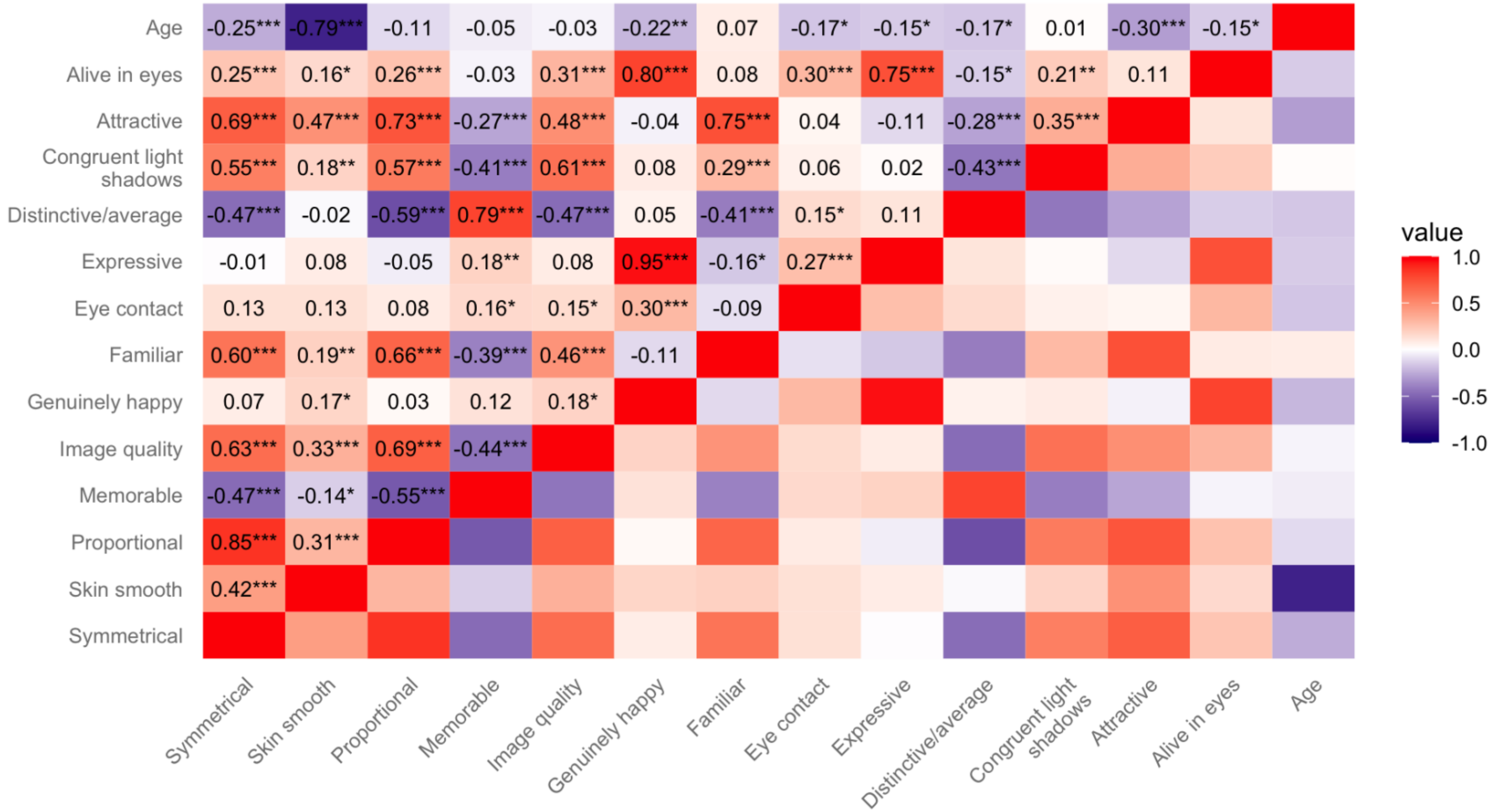
Attribute	β	*SE*	*z*	*p*	95% CI
Proportional	0.55	0.007	6.36	**< .001*****	[0.38, 0.71]
Alive in the eyes	0.28	0.006	4.59	**< .001*****	[0.16, 0.40]
Expressive	0.20	0.005	1.87	0.062†	[-0.01, 0.41]
Familiar	0.17	0.006	2.95	**0.003****	[0.06, 0.29]
Eye contact	-0.01	0.001	-0.16	0.872	[-0.08, 0.07]
Distinctive/average	-0.02	0.006	-0.41	0.682	[-0.14, 0.09]
Image quality	-0.07	0.006	-1.23	0.219	[-0.17, 0.04]
Congruent lighting	-0.08	0.003	-1.78	0.075†	[-0.17, 0.01]
Age	-0.10	0.050	-1.34	0.181	[-0.25, 0.05]
Memorable	-0.13	0.007	-2.02	**0.028***	[-0.25, -0.01]
Symmetrical	-0.17	0.006	-2.49	**0.013***	[-0.31, -0.04]
Attractive	-0.23	0.005	-3.25	**0.001****	[-0.37, -0.09]
Genuinely happy	-0.23	0.060	-1.96	**0.049***	[-0.46, 0.00]
Smooth skinned	-0.48	0.004	-6.20	**< .001*****	[-0.63, -0.33]

Note. Bold text indicates $p < .05$. ***$p < .001$. **$p < .01$. *$p < .05$. †$p < .10$.

Supplement S12 — Correlations Between Face Attributes

Figure S12

Spearman’s ρ Correlations Between Stimulus-Level Mean Ratings Across the 14 Attributes in Experiment 2



Supplement S13 — Lens Model Effects Table

Table S13

Standardized Path Coefficients and Indirect Effects for Each Attribute Explaining Why AI Faces are Misjudged as Human (Ordered by Indirect Effect Size)

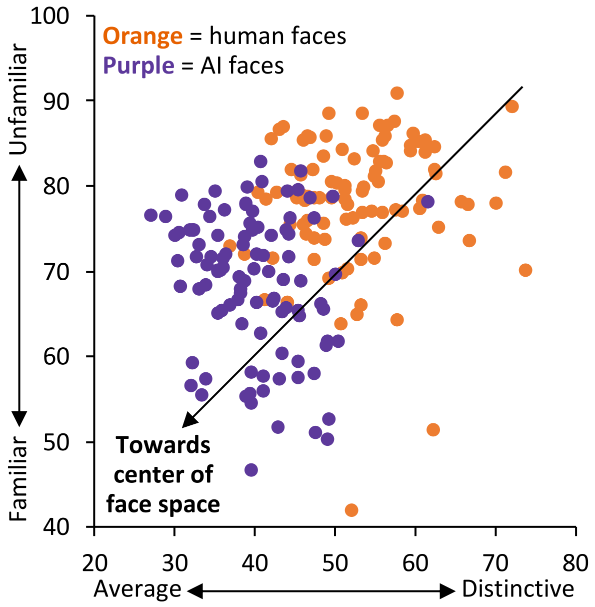
	Standardized coefficients		
Attribute	a path: ass. btw face type & attribute	b path: ass. btw attribute & % judged as human	Indirect effect
Proportional	**-.81*****	**.58*****	**-.47*****
Familiar	**-.52*****	**.21****	**-.11****
Memorable	**.64*****	**-.15***	**-.10***
Alive in the eyes	**-.15***	**.34*****	-.05†
Genuinely happy	.11	**-.31*****	-.03
Eye contact	.12†	.01	~.00
Distinctive/average	**.68*****	.01	~.00
Age	-.04	-.12	~.00
Smooth skinned	-.11	**-.56****	.06
Image quality	**-.60*****	-.10	.06
Expressive	**.19****	**.31***	.06†
Congruent lighting	**-.56*****	**-.12***	**.07***
Symmetrical	**-.68*****	**-.22***	**.15***
Attractive	**-.54*****	**-.28*****	**.15****

Note. a path is the relationship between the face type (human vs AI) and each attribute, with higher coefficients indicating an attribute being exemplified to a greater extent by humans. b path is the relationship between each attribute and the percentage of Experiment 1 participants who judged each face as human. Each indirect effect is the effect of face type on the judgment of the face as human via the specific attribute. Bold text indicates p < .05. ***p < .001. **p < .01. *p < .05. †p < .10.

Supplement S14 — AI and Human Face Averageness

Figure S14

Comparison of AI and Human Faces for Distinctiveness/Averageness and Familiarity Attributes



Note. Familiarity ratings have been reverse scored, so lower scores reflect higher familiarity. Faces are positioned by their mean ratings.

References

Bowling, N. C., & Banissy, M. J. (2017). Emotion expression modulates perception of animacy from faces. *Journal of Experimental Social Psychology*, *71*, 83–95. https://doi.org/10.1016/j.jesp.2017.02.004

Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2018). Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. *Journal of Personality and Social Psychology*, *114*(1), 75–94. https://doi.org/10.1037/pspi0000110

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149. https://doi.org/10.1080/01621459.1987.10478551

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*. 8110–8119. https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html

Khalid, S., Deska, J. C., & Hugenberg, K. (2016). The eyes are the windows to the mind: Direct eye gaze triggers the ascription of others' minds. *Personality and Social Psychology Bulletin*, *42*(12), 1666–1677. https://doi.org/10.1177/0146167216669124

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

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, *119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119

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., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review*, *5*(4), 659–669. https://doi.org/10.3758/BF03208842

Saito, T., Almaraz, S. M., & Hugenberg, K. (2022). Happy = human: A feeling of belonging modulates the "expression-to-mind" effect. *Social Cognition*, *40*(3), 213–227.

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science*, *26*(1), 39–47. https://doi.org/10.1177/0956797614554955

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

Vaitonytė, J., Blomsma, P. A., Alimardani, M., & Louwerse, M. M. (2021). Realism of the face lies in skin and eyes: Evidence from virtual and human agents. *Computers in Human Behavior Reports*, *3*, 100065. https://doi.org/10.1016/j.chbr.2021.100065

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