Face Averageness as a Predictor of Perceived Realism and Trustworthiness in Synthetic Faces

Matthew Ivory
Lancaster University
Lancaster
matthew.ivory@lancaster.ac.uk

Sophie J. Nightingale

Abstract

Human perception of faces is theorized to occur in face-space, where faces are placed within a multi-dimensional space with dense clusters or regions indicating typicality or increased averageness. It has previously been suggested that the increased realism and trustworthiness of synthetic faces may be a result of their increased averageness compared to real faces. In this paper, the averageness of real and synthetic faces (GAN-generated and Diffusion-generated) was calculated using a generalized Procrustean analysis to assess the impact of averageness on perceived realism and trustworthiness. As averageness increased, no meaningful relationship was observed with the correct classification of real, GAN, or Diffusion faces, revealing a uniform accuracy rate across averageness. The relationship between trustworthiness and averageness was modulated by both gender and race, demonstrating complex relationships between real and synthetic faces as well as between the two synthetic generators. The findings highlight the complex ways synthetic faces occupy human face-space and, as such, have important implications for researchers using synthetic faces instead of real ones, particularly in perceptual studies.

1. Introduction

Artificially synthesized faces have been shown to be indistinguishable from real ones and, perhaps due to their averageness, are often perceived as more trustworthy [5, 17]. Others have reported that some synthetic faces exhibit 'hyperrealism' [15] and 'hyperaverageness' [7], with suggestions that the generative process can synthesize faces with features that tap into human perceptual mechanisms, making them appear even more real—and more prototypical—than actual human faces. Critically, previous suggestions that increased facial averageness contributes to increased realism or trustworthiness have been unsubstantiated, particularly from a geomorphic perspective. In this analysis, the geomorphic averageness of real and synthetic faces is examined in relation to human classification accuracy and perceived trustworthiness.

The relationship between trustworthiness, averageness, and other characteristics (e.g., attractiveness) in faces has long been of interest, due to the rich social information that faces provide [9]. In typical social interactions, the face is one of the main ways people decide whether or not to trust another person [18]. Importantly, the increase in perceived trustworthiness of synthetic faces is a relatively new phenomenon that can be linked to technological advancements. Research with earlier versions of face generators reported lower trustworthiness for synthetic faces than real faces [1], whereas generators introduced in the last decade produce faces that have not only become indistinguishable from real faces, but also more trustworthy [17]. With advances in synthetic face generation and their increased usage in online/digital interactions, it is important to understand how these synthetic faces fit into current theories of face perception.

Face-space theory suggests that faces are represented in the brain in a complex multidimensional space, with a centroid that represents a prototypical face. Faces with distinct or asymmetrical features sit further away from the central point [23, 24]. The distribution of faces within this space is not even: theory suggests that more average faces cluster in dense areas, and more distinct faces are found in more sparsely-populated areas. Distinct faces are more easily remembered as they are less likely to be confused with neighboring faces. Faces closer to the centroid will not only be more typical, but perceived as more familiar to a viewer, as they are located closer to the central face (according to the norm-based version of face-space. This paper does not offer a discussion on the two principal models of face-space—

*This work was supported by the Medical Research Council [grant number MR/Y018397/1].

norm-based and exemplar-based—as this is out of scope. For a discussion on the two versions, refer to [23, 24]).

The wider literature suggests that increased familiarity is associated with increased perceived trustworthiness and attractiveness [12, 16, 22, 27]. For real faces, perceived typicality is positively related to trustworthiness, whereas attractiveness does not demonstrate a similar relationship [21]. Given that synthetic faces tend to display lower morphological diversity (higher typicality) than real faces [3], it can be reasoned that they are sampled from, or generated based on, a reduced face-space, or from denser face-space regions. Faces based on these denser regions would appear more typical and therefore more familiar, potentially explaining why synthetic faces are often perceived as more trustworthy than real ones [15, 17]. The idea that synthetic faces are sampled from high-density regions or from a reduced face-space is also supported by complementary findings that super-recognizers of real faces use potential indicators of face-space centrality (such as perceived distinctiveness and perceived symmetry) to classify faces as synthetic resulting in better detection rates than non-super-recognizers [7].

Due to the realism of faces generated with state-of-the-art synthesizers, it is not unreasonable to expect synthetic faces to fall within a face-space of real faces and to be perceived and judged along similar dimensions. To assess how the geomorphic averageness of faces influences perceptions of classification accuracy and trustworthiness, this paper seeks to answer the following questions:

1. Does increased face averageness correlate with classification accuracy

2. Does increased face averageness correlate with perceived trustworthiness?

In this paper, the averageness of faces (determined through its proximity to a prototypical real face) is used as a predictor against classification accuracy as well as perceived trustworthiness. The analyses are based on an existing dataset of realism and trustworthiness ratings for 1,200 images of faces from three sources: real, StyleGAN2-generated, and Diffusion-generated [14]. From this, the present research calculated the proximity of each face from a prototypical real face based on both gender and race groupings before using these values as predictors against realism and trustworthiness.

2. Methodology

2.1. Face Datasets

The real faces were sampled from the Flickr-Faces-HQ Dataset [10] as used in [17]. The StyleGAN2 faces were also taken from [17]. The diffusion faces were generated in Adobe Firefly by [14]. Adobe Firefly is a proprietary model that is trained on Adobe stock and open-licensed images.

Each dataset comprised 400 images balanced for gender (male, female) and race (Black, East Asian, South Asian, and White) with 50 images in each gender-race grouping. Images were manually screened to remove artifacts that were obvious indicators of being synthetic. Images depicted a centered, unobscured head-and-shoulder portrait of a single face. The images had a resolution of $1{,}024^2$ pixels and were aligned by the y coordinates of the eyes.

McGuire et al [14] presented participants with a subset of the 1,200 faces and recorded realism ratings (binary response of 'real' or 'synthetic' when asked to classify each face) and perceived trustworthiness on a scale of 1-7 (very untrustworthy to very trustworthy).

GANs consist of two adversarial networks: a generator that maps samples from a low dimensional latent vector to images, and a discriminator that learns to distinguish generated images from real ones. Training proceeds as a minimax game in which the generator is penalized whenever the discriminator correctly identifies a synthetic image as fake; over time the generator learns to map dense regions of its latent space to high quality, plausible faces [11, 25]. Faces generated closer to this mean center point are likely to be more symmetrical and possess fewer artifacts, whereas faces produced outside these dense areas can result in atypical images and reflect a departure from the model's high-density, best-learned region.

Diffusion models generate images by learning a stochastic denoising process: during training, progressively increasing noise is added to the images and the model learns the reverse (denoising) Markov process that maps noise back to data [19]. When generating images, the model aims to denoise random Gaussian noise towards a data sample (such as text prompts). Unlike the canonical GAN latent vector, some diffusion architectures operate directly in pixel space and others (e.g., latent diffusion variants) perform denoising in a compressed latent representation. Diffusion models are based on a fundamentally different generative mechanism than GAN models [4]. Empirically, diffusion models often exhibit greater diversity and more consistent coverage of the training images (fewer extreme artifacts), while GANs are more prone to limited diversity via mode collapse [8] or truncation [10], differences that may plausibly influence where the generated faces fall in a geomorphic face-space affecting averageness measures.

2.2. Calculating Face Averageness

Four hundred real faces (taken from the Flickr-Faces-HQ dataset, as used by Nightingale and Farid [17]) were centered, scaled by interocular distance, and rotated, before being decomposed into their 478 unique facial landmarks, as determined via Mediapipe's face landmark detection algorithm. The mean average coordinates (x, y, z) for these landmarks were calculated. The use of a Procrustean coordinate

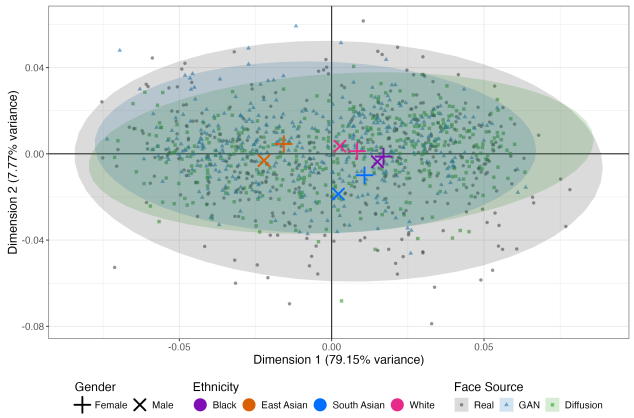

Figure 1. Distribution of faces in principal component space. Each point represents a single face, projected onto the first two principal components of Procrustes-aligned landmark coordinates. Larger symbols indicate the real face centroids. Shaded regions show 95% confidence regions derived from each face source's covariance matrix, illustrating the spatial overlap and relative dispersion.

system ensured all faces were aligned in a shared coordinate system [26]. From this, the images were separated based on gender and race, creating eight different groups from male and female, as well as Black, East Asian, South Asian, and White faces. Each of these groups was used to create the prototypical face to which all other images of the same groupings were compared, culminating in eight prototypical faces. We then conducted the same Procrustean analysis on all faces and computed each face's Euclidean distance from their respective prototypical face, as the square root of the sum of squared distances across each landmark [27]. This resulted in a single positive value for each face with values tending towards zero indicating greater proximity to the mean face (and thus being 'more average'). Values were multiplied by -1 so that higher values indicated greater averageness.

2.2.1. Data Analysis

Procrustean analysis was carried out with Python and the resulting dataset was manipulated and analyzed in R. Linear regression models were used to model both accuracy and trustworthiness ratings. Linear regressions were compared against Generalized Additive Models (GAM) to account for the curvilinear nature of the relationships, but in both cases, the linear model was of a similar or better fit. Averageness scores were scaled and transformed into standard deviation units for interpretability. The relationship of the dependent variables were explored through trendline analyses to understand how changes in accuracy or trustworthiness are modeled.

3. Results

The overall averageness (unscaled) for real faces was *M* = -0.053 (*SD* = 0.014), for GAN faces it was *M* = -0.046 (*SD* = 0.013) and for diffusion faces it was *M* = -0.047 (*SD* = 0.011), indicating slightly increased averageness for the synthetic faces. This can be seen in Figure 1 where the distance of individual faces is visualized as the principal components of Procrustes shape coordinates of the faces. The synthetic faces are predominantly positioned within the real face-space, supporting the idea that synthetic faces occupy a more densely-populated region or reduced face-space than real faces [3]. Figure 2 shows examples of the four most average and four least average faces.

Pairwise comparisons, with Tukey adjusted p-values, between these face sources indicates a non-significant difference of 0.0009 between the averageness (unscaled) of the two synthetic sources, $t(1197) = 0.99$, $p = 0.585$. A significant difference of 0.0067 was observed between GAN and real images, $t(1197) = 7.27$, $p < 0.001$ indicating that GAN images are more average overall, with a similar finding for Diffusion and real faces, with a mean difference of 0.0058, $t(1197) = 6.28$, $p < 0.001$. This lends further support to synthetic images occupying a reduced face-space compared to real faces.

Table 1. Estimated Accuracy Trends for Averageness by Source, Gender, and Race.

Source	Gender	Race	Trend	p
Real	Female	Black	−0.03	0.264
		East Asian	0.00	0.919
		South Asian	0.00	0.897
		White	0.01	0.680
	Male	Black	0.05	0.053
		East Asian	−0.01	0.817
		South Asian	0.03	0.332
		White	0.01	0.776
GAN	Female	Black	−0.05	0.081
		East Asian	−0.04	0.347
		South Asian	−0.01	0.612
		White	−0.03	0.264
	Male	Black	−0.01	0.537
		East Asian	−0.03	0.229
		South Asian	−0.04	0.231
		White	−0.03	0.269
Diffusion	Female	Black	−0.03	0.383
		East Asian	−0.01	0.696
		South Asian	0.01	0.812
		White	−0.02	0.610
	Male	Black	0.03	0.409
		East Asian	−0.04	0.127
		South Asian	−0.04	0.242
		White	−0.03	0.326

Figure 2. The top row shows the most average faces from L-R with scaled averageness values: real White male (2.31), GAN South Asian female (2.05), GAN South Asian male 1.97), and GAN White female (1.95). The bottom row shows the least average faces with (L-R): real Black male (-3.43), real South Asian male (-3.46), real East Asian female (-3.55), and Diffusion East Asian male (-3.56).

3.1. Accuracy and Averageness

To address the first research question, *does increased face averageness correlate with classification accuracy*, a linear model was used to test predictors of scaled averageness, face source, gender, and race, along with the four-way interaction between these, against the accuracy ratings for each face. Here, accuracy refers to the proportion of correct responses made for each image as a value between 0 and 1. Compared to a GAM, AIC scores were identical (-424.98) and so was the R^2_{adj} value (0.23). For ease of interpretation, the linear model was chosen for further analysis. The linear model was significant, $F(47, 1152) = 8.50$, $p < 0.001$. All significance tests of trends are shown in Table 1.

A trendline analysis tested whether accuracy changed as averageness increased. For all combinations of race and gender, there was no significant change in accuracy as averageness increased, presenting a uniform relationship, see Figure 3. As a result, no correlation was observed between increased face averageness and classification accuracy when accounting for variations of gender and race.

3.2. Perceived Trustworthiness and Averageness

To answer the research question, *does increased face averageness correlate with perceived trustworthiness*, trustworthiness was modeled through predictors of scaled averageness, face source, gender, and race along with the interaction between these four predictors. The model was significant, $F(47, 1152) = 18.36$, $p < 0.001$ with $R^2_{adj} = 0.41$. Predictors of facial yaw, gradient detection, and luminance did not improve the model significantly and were omitted.

Due to the complex four-way interaction, the remainder of the results report only the significant findings in-text. All significance tests are shown in Table 2 and the relationship between trustworthiness and averageness is shown in Figure 4.

A trendline analysis was conducted for all eight groups of gender and race. Within the real faces, Black female faces significantly increased in trustworthiness by a rate of 0.23, $t = 2.65$, $p < 0.001$, a similar trend to Black male faces, $\beta = 0.38$, $t = 4.25$, $p < 0.001$. Whereas East Asian male faces decreased in trustworthiness, $\beta = -0.21$, $t = -2.21$, $p = 0.028$. For all other faces, no significant trends were seen, indicating uniform trustworthiness over averageness.

For GAN faces, five of the eight face types significantly increased in trustworthiness as averageness increased. This included female South Asian faces, $\beta = 0.37$, $t = 3.81$, $p < 0.001$, and White faces, $\beta = 0.66$, $t = 5.83$, $p < 0.001$. Within male GAN types, Black faces significantly increased, $\beta = 0.40$, $t = 4.48$, $p < 0.001$, as did South Asian faces, $\beta = 0.48$, $t = 4.19$, $p < 0.001$, and White faces, $\beta = 0.24$, $t = 2.22$, $p = 0.027$.

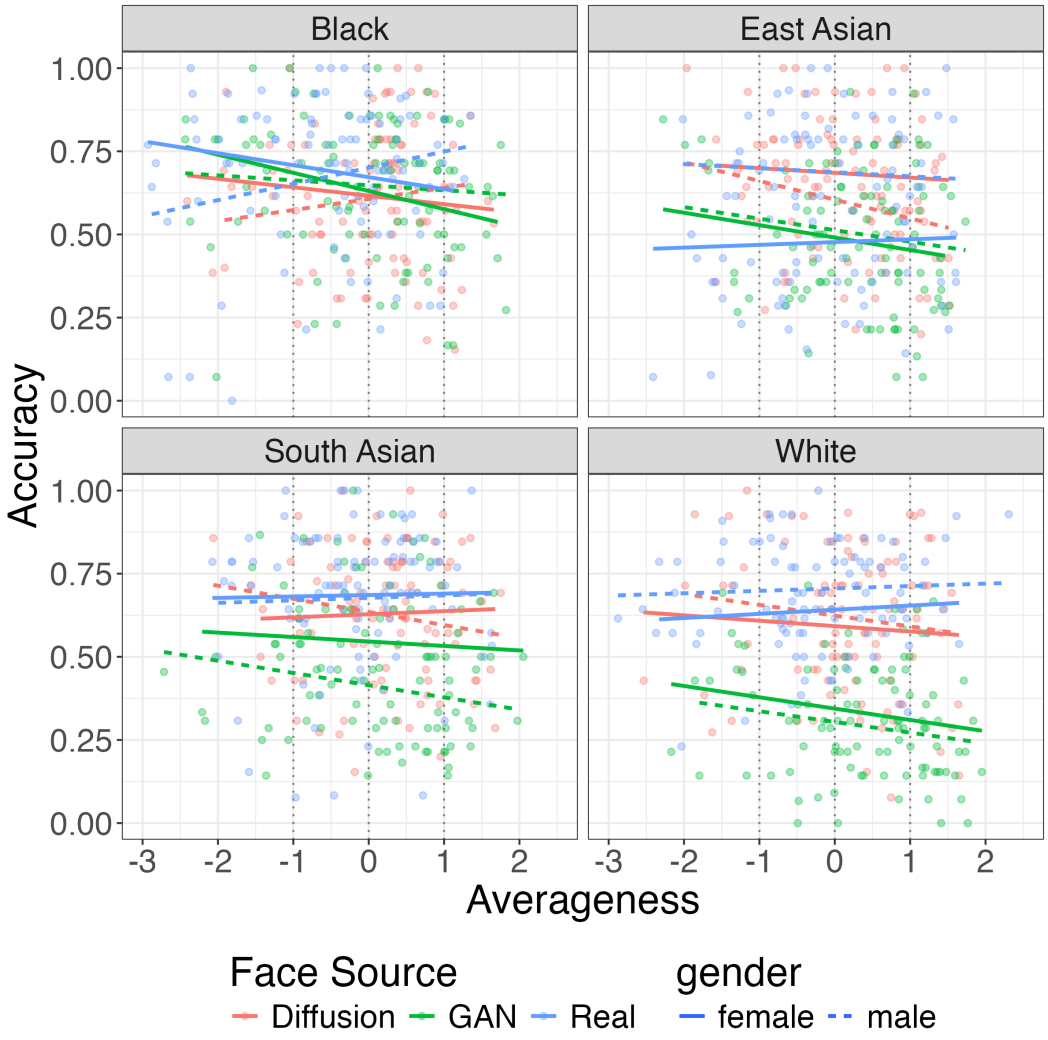

Figure 3. The accuracy-averageness relationships across the three face sources separated by race and gender. Vertical dotted lines indicate the three SD points.

For Diffusion faces, female Black faces were also significantly linked to increases in trustworthiness, $\beta = 0.25$, $t = 2.37$, $p = 0.018$, as were South Asian faces, $\beta = 0.29$, $t = 2.03$, $p = 0.043$. Within the male grouping, only Black faces were significant, $\beta = 0.43$, $t = 2.86$, $p = 0.004$.

Taken together, apart from female Black GAN images, all Black faces significantly increased in perceived trustworthiness as their geomorphic averageness tended towards their respective centroid. Within the synthetic faces, South Asian faces also positively trended for both GAN and Diffusion female images, but only for male GAN faces, not Diffusion. Only GAN images showed a significant positive trend for white faces, with female faces having the largest increase in trustworthiness per standard deviation increase of averageness. Of all the significant trends, only real East Asian male faces showed a decrease in perceived trustworthiness as averageness increased.

4. Discussion

A generalized Procrustean analysis was used to compute averageness scores for 1,200 faces (real, StyleGAN2-generated, and Diffusion-generated). These scores were analyzed alongside prior data on classification accuracy (real or synthetic) and perceived trustworthiness. No significant trend was seen between accuracy and averageness, and the relationship between trustworthiness and averageness was complex, with interactions between gender and race.

4.1. Realism and Averageness

In answering the first research question, 'does increased face averageness correlate with classification accuracy', an

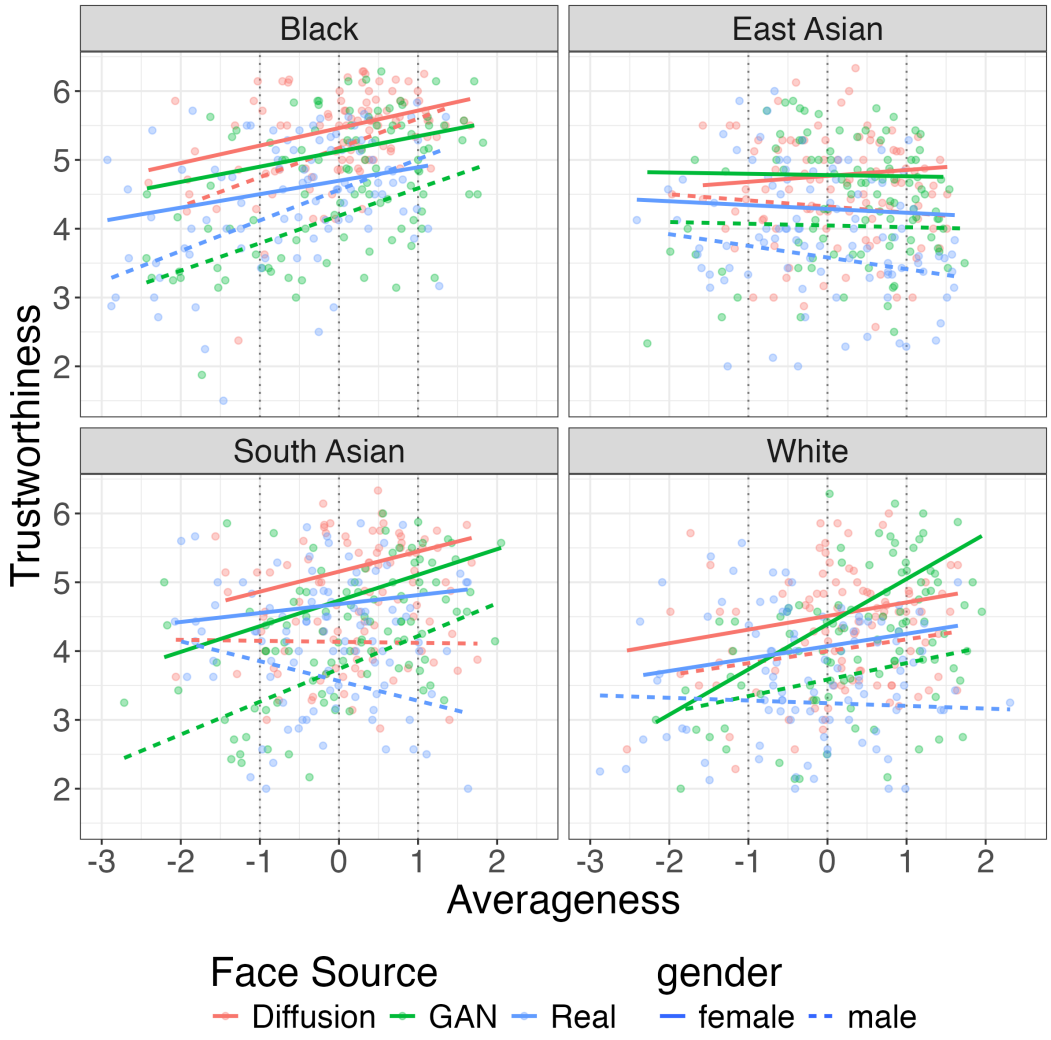

Figure 4. The trustworthiness-averageness relationships across the three face sources separated by race and gender. Vertical dotted lines indicate the three SD points.

answer is provided: 'no significant trend was seen'. As averageness increased, accuracy was uniform within groupings of gender and race, indicating averageness had minimal influence on accuracy. Previous studies suggested that the increased averageness of synthetic faces may be a contributing factor of their increased realism [15, 17]. The present findings reveal that synthetic faces are significantly more geomorphically average than real faces overall (as indicated by pairwise comparisons between the averageness of the three face sources), however, as seen in Figure 3 and Table 1, an increase in averageness within each face source did not change perceived realism (which would have been observed through decreased accuracy levels for synthetic faces as averageness increased). Previous work has reported that super-recognizers use perceived averageness for classifying faces with some success [7], but the present findings did not find this for statistical averageness. Therefore it is possible that statistical and perceptual averageness are separate entities that do not closely correspond, or that super-recognizers are relying on a different cue that they mistakenly interpret as averageness.

4.2. Trustworthiness and Averageness

In answering the second question, 'does increased face averageness correlate with perceived trustworthiness', the answer is complex, with the intersection of gender and race highlighting that the averageness and trustworthiness relationship is not consistent across gender or race. From a total of 24 combinations, 13 of the combinations of face source, gender, and race demonstrated no significant change in trustworthiness as averageness increased. Eleven of the combinations did reveal significant effects, with some clear

Table 2. Estimated Trustworthiness Trends for Averageness by Source, Gender, and Race. Significant trends at the 0.05 threshold are in bold.

Source	Gender	Race	Trend	p
Real	Female	**Black**	**0.23**	**0.008**
		East Asian	−0.01	0.899
		South Asian	0.13	0.291
		White	0.18	0.109
	Male	**Black**	**0.38**	**<0.001**
		East Asian	**−0.21**	**0.028**
		South Asian	−0.14	0.139
		White	−0.04	0.679
GAN	Female	Black	0.22	0.056
		East Asian	−0.02	0.902
		South Asian	**0.37**	**<0.001**
		White	**0.66**	**<0.001**
	Male	**Black**	**0.40**	**<0.001**
		East Asian	−0.03	0.814
		South Asian	**0.48**	**<0.001**
		White	**0.24**	**0.027**
Diffusion	Female	**Black**	**0.25**	**0.018**
		East Asian	0.09	0.524
		South Asian	**0.29**	**0.043**
		White	0.20	0.093
	Male	**Black**	**0.43**	**0.004**
		East Asian	−0.06	0.563
		South Asian	−0.01	0.906
		White	0.17	0.157

trends emerging both across and within face source, gender, and race.

Across the three face sources (real, GAN, and Diffusion), ratings of trustworthiness increased as averageness increased for Black faces, apart from for female Black GAN faces (where there was a non-significant positive relationship that tended towards significance). For both male and female faces, trustworthiness increased for South Asian and White GAN faces, as well as for female South Asian Diffusion faces. The only instance in which increased averageness was associated with decreased ratings of trustworthiness was for East Asian real male faces. All other interactions of gender and race were non-significant, representing more uniform relationships between trustworthiness and averageness. Why then, is a positive relationship between trustworthiness and averageness observed for some, but not all groups of images? A plausible, though speculative, explanation for this might be that there is an interaction between perceptual (human) and technological mechanisms.

Trustworthiness judgments are known to draw on a range of facial cues, including averageness [21], attractiveness [13], happiness [18], and femininity [6]. As well as these established cues that influence judgments of trust, the current findings also reveal that whether a face is real or artificially-synthesized also impacts our perceptions. That said, our trustworthiness model explained only around 41% of the data variance, suggesting that trustworthiness is derived from more than just averageness, face source, gender, and race. Of course, our model did not include cues such as happiness and femininity (beyond geomorphic averageness of female faces) and thus the residual variance may be attributable to human perception of such cues, or to technical properties of the images themselves, including GAN-related resampling artifacts (e.g., the truncation trick). Trustworthiness is likely further modulated by individual differences across the human raters, for example: own-race faces tend to be rated as more trustworthy [20], and women appear to use different cues depending on the direction of their judgment—placing more reliance on averageness when judging faces as less trustworthy and more reliance on attractiveness when judging faces as more trustworthy [13].

The differences in the elevated trustworthiness of some GAN faces may be, in part, an illusion engineered by the model. The truncation trick [10] resamples low-probability regions of the latent space towards higher-probability ones, enhancing image quality at the cost of reduced variety. In other words, the truncation trick shrinks the sampling range of the latent space around the mean, so that only feature combinations close to the average are drawn upon when generating an image. The consequence is that GAN faces may be statistically forced to be average, constrained to a reduced region of feature space with no direct parallel in the natural distribution of real faces. Faces occupying these high-density latent regions carry the same feature signatures as genuinely average human faces, hijacking the mechanisms that ordinarily guide trustworthiness judgments. Since facial averageness is a well-established correlate of perceived trustworthiness [18], GAN faces may inflate trustworthiness ratings not because they possess genuine social cues of trustworthiness, but because they have been artificially compressed into the region of face-space that the perceptual system treats as trustworthy. For example, GAN resampling may result in faces with features that look perceptually more trustworthy when combined with other characteristics such as skin tone. This may result in faces that are not true representations of real faces as they are a 'combination' of many different races. GAN averageness could therefore be considered a statistical artifact that deceives perceptual judgments. These statistical artifacts may be more deceptive in combination with specific gender and race groupings, such as for White or South Asian GAN faces.

In comparison, Diffusion models do not apply the truncation trick; instead, by learning the full distribution of facial features through the noise-denoising process [4], they generate diverse faces whose trustworthiness ratings are

grounded in plausible human variation rather than compression into an artificially typical region of latent space, which may explain the weaker or reduced number of significant relationships between averageness and trustworthiness within Diffusion faces.

Given the complex interactions in the model, it is more appropriate to acknowledge this complexity rather than over-interpret the individual effects. The pattern of results for White GAN faces could partially support both the hyperrealism effect [15] and GAN hyperaverageness [7], however, this does not explain why a similar pattern for South Asian GAN faces was seen. Overall, the interplay of gender, race, face source, and averageness is nuanced and requires careful consideration when using these images in practice.

4.3. Implications

The finding that increasingly average faces were rated as more trustworthy, particularly South Asian and White GAN faces, suggests that some faces could pose greater risk when used for deceptive purposes. Nonetheless, given that many gender and race groupings showed no trustworthiness differences across levels of averageness (e.g., synthetic East Asian faces), it may be the case that malicious actors need not be highly selective in choosing faces. Therefore, synthetic faces offer a potentially powerful tool for social engineering/psychological manipulation, though their effectiveness varies by demographic and generation method.

The findings from this study have meaningful implications for research going forward. If researchers choose to use synthetic faces, particularly from GAN generators, they should be aware of the impact that averageness can have on human perception of the images. Although synthetic faces have been widely used and recommended as experimental stimuli [2], the present findings suggest that they cannot be treated as a simple replacement merely because they are easier to create and manipulate—this is especially the case when such images are to be used in research involving human perceptual judgements and ratings. Differences in averageness of GAN faces resulted in significant changes in trustworthiness compared to real faces across combinations of gender and race, and so without adequate controlling of the stimuli the examined effects may be modulated or attenuated by facial averageness. Without adequate consideration of the synthetic stimuli, researchers could find that the results are influenced by properties seemingly unique to synthetic faces, particularly GAN images.

One limitation of this research was that face averageness was taken from head-and-shoulder images of faces complete with backgrounds. The facemesh used to isolate facial characteristics removes all information outside of this, including ears, hair, clothing, and the background, and so whilst the trustworthiness model indicated reasonable statistical explanation, (R^2_{adj} = .41), it may underemphasize the role of the non-face content within the images. The original research paradigm presented the complete $1{,}024^2$ size images to individuals and asked them to judge the faces as real or synthetic, however it was unknown how much non-face information would have been incorporated into the decision making process. Future work should explore synthetic face detection with all non-face information removed to determine how this influences detection rates, as this may influence trustworthiness ratings of synthetic imagery. Comparing perceived averageness with geomorphic averageness could be beneficial for determining whether people can detect increased statistical averageness within synthetic faces.

Additionally, this study used a subset of images generated by two synthetic sources and is not necessarily representative of all GAN and Diffusion models. Despite this, the findings evidence the existence of differences in averageness of synthetic faces and real faces. Further work should incorporate larger numbers of images and generators to see whether the effects are model specific, or apply more widely to architectures.

4.4. Conclusion

The geomorphic averageness of three face sources: real, GAN, and Diffusion, was assessed against perceived realism and trustworthiness. No significant relationship was observed between averageness and realism (as measured by classification accuracy) for real, GAN, or Diffusion faces. The relationship between averageness and trustworthiness was complex; although 10 of the 24 comparisons revealed a positive relationship, this finding was not universal with differences across the groupings of face source, gender and race. GAN faces showed the strongest effect of averageness on trustworthiness, particularly for South Asian and White faces. In comparison, real and Diffusion faces were more likely to present monotonic ratings of perceived trustworthiness across averageness. Ultimately, statistical averageness alone cannot account for the perceptual realism and trustworthiness of synthetic faces, highlighting the need for further research. The positioning of synthetic faces within human face-space offers important insights for researchers considering them as alternatives to real faces.

References

[1] Benjamin Balas and Jonathan Pacella. Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior*, 77:240–248, 2017. 2

[2] Casey Becker, Russell Conduit, Philippe A. Chouinard, and Robin Laycock. Can deepfakes be used to study emotion perception? A comparison of dynamic face stimuli. *Behavior Research Methods*, 56(7):7674–7690, 2024. 9

[3] Olga Boudníková and Karel Kleisner. AI-generated faces show lower morphological diversity than real faces do. *Anthropological Review*, 87(1):81–91, 2024. 3, 4

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 3, 8

[5] Alexander Diel, Tania Lalgi, Isabel Carolin Schröter, Karl F. MacDorman, Martin Teufel, and Alexander Bäuerle. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16:100538, 2024. 2

[6] Yan Dong, Yi Liu, Yanfei Jia, Yongna Li, and Chen Li. Effects of Facial Expression and Facial Gender on Judgment of Trustworthiness: The Modulating Effect of Cooperative and Competitive Settings. *Frontiers in Psychology*, 9:2022, 2018. 8

[7] James D. Dunn, David White, Clare A. M. Sutherland, Elizabeth J. Miller, Ben A. Steward, and Amy Dawel. Too good to be true: Synthetic AI faces are more average than real faces and super-recognizers know it. *British Journal of Psychology*, 00:1–16, 2026. 2, 3, 7, 9

[8] Yanxiang Gong, Zhiwei Xie, Guozhen Duan, Zheng Ma, and Mei Xie. Distribution Fitting for Combating Mode Collapse in Generative Adversarial Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):18251–18262, 2024. 3

[9] Rachael E. Jack and Philippe G. Schyns. The Human Face as a Dynamic Tool for Social Communication. *Current Biology*, 25(14):R621–R634, 2015. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3, 8

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. pages 8110–8119, 2020. 3

[12] Karel Kleisner, Zuzana Štěrbová, and Vojtěch Fiala. What constitutes the perception of facial typicality? *Personality and Individual Differences*, 231:112838, 2024. 3

[13] Nan Li and Ning Liu. The Nonlinear and Gender-Related Relationships of Face Attractiveness and Typicality With Perceived Trustworthiness. *Frontiers in Psychology*, 12, 2021. 8

[14] Alexis McGuire, Matyas Bohacek, Hany Farid, Paul Taylor, and Sophie Nightingale. AI-generated faces are becoming more trustworthy (preprint). *OSF*, 2026. 3

[15] Elizabeth J. Miller, Ben A. Steward, Zak Witkower, Clare A. M. Sutherland, Eva G. Krumhuber, and Amy Dawel. AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, 34(12):1390–1403, 2023. 2, 3, 7, 9

[16] Tamami Nakano and Takuto Yamamoto. You trust a face like yours. *Humanities and Social Sciences Communications*, 9 (1):226, 2022. 3

[17] Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8): e2120481119, 2022. 2, 3, 7

[18] Nikolaas N. Oosterhof and Alexander Todorov. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092, 2008. 2, 8

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[20] Irina Schmid, Zachary Witkower, Friedrich M. Götz, and Stefan Stieger. Registered report: Social face evaluation: Ethnicity-specific differences in the judgement of trustworthiness of faces and facial parts. *Scientific Reports*, 12(1): 18311, 2022. 8

[21] Carmel Sofer, Ron Dotsch, Daniel H. J. Wigboldus, and Alexander Todorov. What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science*, 26(1):39–47, 2015. 3, 8

[22] Alexander Todorov, Manish Pakrashi, and Nikolaas Oosterhof. Evaluating Faces on Trustworthiness After Minimal Time Exposure. https://guilfordjournals.com/doi/10.1521/soco.2009.27.6.813, 2009. 3

[23] Tim Valentine. 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, 1991. 2, 3

[24] Tim Valentine, Michael B. Lewis, and Peter J. Hills. Face-Space: A Unifying Concept in Face Recognition Research. *Quarterly Journal of Experimental Psychology*, 69(10): 1996–2019, 2016. 2, 3

[25] Xin Wang, Ting Yu Tsai, Li Lin, Hui Guo, Shu Hu, Ming-Ching Chang, Pradeep K. Atrey, and Siwei Lyu. Spotting the Fakes: A Deep Dive into GAN-Generated Face Detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 21 (7):193:1–193:24, 2025. 3

[26] Miriam Zelditch, Donald Swiderski, H. David Sheets, William L. Fink, Miriam Zelditch, Donald Swiderski, H. David Sheets, and William L. Fink. *Geometric Morphometrics for Biologists: A Primer*. Elsevier Science & Technology, Chantilly, UNITED STATES, 2004. 4

[27] Amy A. Z. Zhao, Keagan Harrison, Alexander Holland, Henry M. Wainwright, Jo-Maree Ceccato, Morgan J. Sidari, Anthony J. Lee, and Brendan P. Zietsch. Objectively measured facial traits predict in-person evaluations of facial attractiveness and prosociality in speed-dating partners. *Evolution and Human Behavior*, 44(4):315–323, 2023. 3, 4