Title

Discriminating real from A.I.-generated faces: Effects of emotion, gender, and age.

Permalink

https://escholarship.org/uc/item/49v4z44k

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Duffy, Sean
August, Antoine
Wisniewski, Kate

Publication Date

2024

Copyright Information

Discriminating real from AI-generated faces: Examining the effects of stimulus age, gender, and emotional expression.

Sean Duffy (seduffy@scarletmail.rutgers.edu)
Department of Psychology, Rutgers University - Camden
Camden, New Jersey, 08102

Antoine Auguste (antoine.g.auguste@gmail.com)
Rutgers University - Camden
Camden, New Jersey, 08102

Kate Wisniewski (wisniewskk@chop.edu)
Children's Hospital of the University of Pennsylvania
Philadelphia, Pennsylvania 19104

Abstract

This paper reports two studies examining participants' identification accuracy in discriminating real faces from realistic "artificial" faces created through the Artificial Intelligence (AI) system StyleGAN. Across the two studies, two different sets of participants (N = 400) attempted to distinguish 24 real from 24 AI-generated images. Both sets of participants exhibited poor discrimination accuracy and a bias to report all images as real (Study 2). We examined other possible influencing factors were examined, such as smile intensity (Study 1) and age-congruence between participants and faces (Study 2). Implications for future research, and for understanding the potential societal impacts of AI-generated online content are discussed.

Keywords: Artificial intelligence, Machine learning, Facial perception

Introduction

In the past few decades, the topic of 'artificial intelligence' (AI) has received considerable attention in the cognitive sciences. (Turing, 1950; Levesque, 2017; Neufeld & Finnestad, 2020; Hsu, et al., 2018; Tariq, et al., 2018; Tolosana, et al., 2020). However, the past two years have witnessed an surge of interest in AI due to the release of several generative programs such as ChatGPT that provide users with new tools that are revolutionizing fields across a wide spectrum of academic disciplines. One question of considerable interest to those employing this technology the quality of the output of generative AI. Indeed, if AI may be used from any application from creating digital content (Israel & Amer, 2023) to screening radiographs for neoplasms (McKinney, et al. 2020), it is necessary to explore the accuracy and veracity of the output of AI. The current study presents two experiments examining whether human participants show the ability to discriminate between real images of human faces from artificially generated faces using an AI implementation known as StyleGAN (Karpathy, et al., 2016).

Background Literature and Rationale

Humans are natural experts at perceiving faces, an ability that has evolved in our species and becomes honed during processes of development (e.g., Farah et al., 1998; Mondloch et al., 1999; Scott & Monesson 2009). This perception is highly accurate and pervasive: Adults may be able to recognize fellow friends after an absence of many decades and natural changes that occur through aging. No doubt this capacity emerges due to its ubiquity, necessity, and importance for social processes such as kin recognition or social advancement.

Until quite recently, the capacity for computers to create artificial stimuli has been quite primitive and rudimentary, no doubt due to the complexity of creating facial stimuli advanced enough to fool the neural system governing facial perception. For example, while it may be possible for a more primitive program to generate abstract stimuli or more concrete stimuli of natural objects such as flowers or cat faces,

Prior to about 2015, while programs existed that produced facial stimuli that closely resembled human faces, human perceivers could still recognize artificial faces from real faces. However, in 2018, a generative adversarial network (GAN) program called StyleGAN (based on Nvidia's CUDA software along with Google's TensorFlow) was released that could generate novel stimuli after learning from a training set of exemplars (for technical information see Karras, et al. (2019; 2021).

The issue of how realistic artificial images of faces appear is relevant because humans exhibit a particular pattern of facial perception whereby faces that are close to exhibiting naturalistic features but fall significantly short elicit negative emotions and evaluations, known as the "uncanny valley." Often discussed in the literature on robotics, the uncanny valley is a region of an evaluative dimension in which human-like facial stimuli appear to be disturbing, atypical, or off-putting because the stimulus is close to appearing human but is missing some key features that all human faces have in common.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

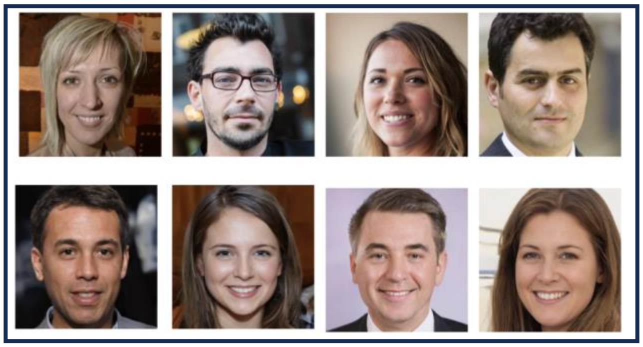

Figure 1. The top row of photos is of real people that were input into the GAN, collected initially from FLICKR. The bottom row is artificially generated faces. The faces in the bottom are created utilizing artificial intelligence.

Figure 1 provides sample images of the real faces and those artificial faces produced by StyleGAN. What is striking is how realistic the artificial images (bottom row) are relative to those that are real photographs. This example demonstrates just how advanced this technology has become and is only improving: these images are already almost five years old and produced with the first version of StyleGAN, which has since gone through two major revisions.

The facial stimuli produced by StyleGAN received considerable media attention in 2019 when the psychologist Carl Bergstrom and biologist Jevin West created a website associated with their popular science book Calling Bullshit in which they invited the public to view side-by-side images of faces – one real from the training set and one produced by StyleGAN – and the viewer's task was to click on the image that featured a "real" person. Since 2019, this website has been visited tens of millions of times (Bergstrom & West, 2022).

While participants were unaware of this, Bergstrom and West (2022) were collecting accuracy data reaction time data. With respect to accuracy, they found that participants were generally above average at discriminating real from artificial faces, with an accuracy rate of over 65 percent, significantly greater than chance (50%). They also found that accuracy increased over time for those participants who iteratively completed and repeated the task over time, suggesting that discrimination performance increased, suggesting learning over trials.

However, a look at the actual StyleGAN output suggests that participants may have been using features of the images unrelated to the faces themselves to establish their veracity. For example, many artificial images contained splotches or "watermarks" that looked like unusual stains that one would rarely find in a digital image. While the origin of these watermarks is unclear, they seem to represent distortions based on highlights or reflections within the images themselves. A second type of cue participants may have learned is by examining the backgrounds of the images. While a face may be rendered perfectly, information in the background that was not learned by the program because there were not enough training examples ended up distorted. So trees or telephone poles might not resemble those one might find in the real world. A third feature is color. Some of the images contained hair exhibiting unnatural colors that anyone familiar with the way dyed hair looks would know was unrealistic. A final issue concerned symmetry, particularly with respect to the mouth. In many of the artificial images, teeth were misaligned with one tooth offset in an unnatural way. So, while the program produced incredible renditions of faces, the products were just abnormal enough to raise flags with participants and may partially explain why performance was well above chance.

There are other issues that are not presently understood about generated images. For instance, it is not known if gender, age, or emotional expression affect ratings of whether the facial stimulus is real (Montagne, et al. 2005; Olderbak, et al., 2019). The present study examines these possible effects.

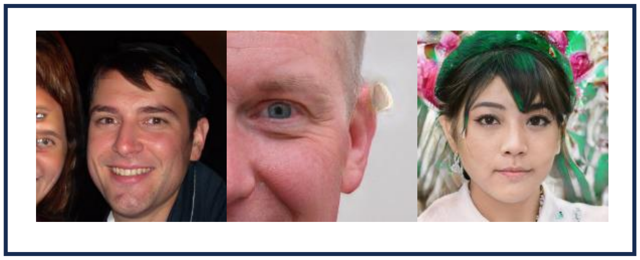

Figure 2: Example of distorted images. Note the third eye in the person on the left, the watermark over the ear in the center image, and the unusual rendering of dyed hair in the right image.

The Studies

Study 1 examined the effect of smile strength on accuracy in discrimination between artificial and real images. Prior research identifies "Duchenne" smiles as genuine and typically wider, with cheek-riser activation. Gunnery and Ruben (2016) conducted a meta-analysis examining perceptions of Duchenne and non-Duchenne smiles. They found that people producing Duchenne smiles are perceived more positively than those producing non-Duchenne smiles. Recent research has found that a Duchenne smile is more of an artifact of smile intensity, rather than an indication of genuine positive emotion (Girard et al., 2019). For artificially created images, does smile strength (similarity to Duchenne smiles) affect perception? The present study categorized and evenly distributed the images presented in the survey across smile strengths ranging from ratings of 0 to 100. These ratings were provided by an application embedded within the popular image sharing software Instagram that rates the extent of people's smile determined by an algorithm (see Arias et al., 2018). Based on prior literature, smiles scoring closer to 100 (more consistent with Duchenne features) were expected to be rated as more authentic. We hypothesized that images showing faces with a stronger smile would be both rated as realer (on a continuous scale) and rated as real more frequently across participants (considering real vs. artificial as binned response categories), than those showing faces that have a weaker smile, regardless of authenticity.

Study 1

Methodology The participants in Study 1 were N = 132 undergraduate students at a eastern state university (n_{Female}=103) who completed an online Qualtrics survey in exchange for partial course credit.

Materials and Design: Participants were presented with 24 images of real faces interspersed with 24 images of artificially generated faces produced by the StyleGAN technology (Karras et al., 2019), for a total of 48 trials. The stimuli used in the study were evenly matched across subsets (real vs. fake) for smile strength, age, and gender. The display order of the faces was randomized across participants. During each trial, participants saw one face and were answered a series of items to rate each image on a 7-point Likert scale. First, participants rated their confidence in whether or not the face they were viewing was of a real person or was AI-generated. The response options ranged from (1) Definitely Artificial to (7) Definitely Real, with option (4) serving as an “Unsure” middle point. Next, participants answered three additional questions regarding each image’s trustworthiness, attractiveness, and usualness. The purpose of including these questions was two-fold:(a) to obscure the purpose of the study; and (b) to evaluate correlates of the perceived trustworthiness, attractiveness, and usualness of faces without specific *a priori* hypotheses. After all trials were completed, participants were asked basic demographic questions, including their age, gender, race, and highest level of education attained. No identifiable information was collected.

Stimuli were collected from the publicly available StyleGAN stimulus repository (Karras et al., 2019). This repository contains 80,000 photos of real faces, originally collected through Flickr (a website used for posting pictures) under a creative commons license, and 100,000 photos of artificially generated faces created through StyleGAN (Karras et al., 2019).

For this study, strict exclusion criteria were established before examining the pre-existing dataset. These exclusion criteria were based on those outlined by Bergstrom & West (2022), with additions. These criteria were established to ensure that the images used in this study were of the highest quality, lacking those features that were cues to being artificially generated. These exclusion criteria were utilized precisely to simplify the process of matching the stimuli sets of real and artificial faces (for smile strength, age, gender, and race). Photos of people with “unnatural” features were excluded, such as individuals with vivid hair colors or dressed in a costume. Another exclusion criterion, typically seen in authentic photos, was low image quality: any images containing lens flare, suboptimal lighting, and/or were out of focus were not included as stimuli in this study.

Specifically with respect to artificially generated images, obvious distortions within the image that were close to the face or distracting to the viewer were excluded. These often appeared as "water spots" mentioned previously that are a clear signal that a photo has been artificially created. Finally, a primary exclusion criterion for all photo stimuli was that any photos containing multiple individuals were not included such as cropped group photos (a frequent occurrence in Flickr images), nor photos in which other people could be seen in focus or close to the main subject. This was primarily done to eliminate facial confounds and irrelevant distractors from the stimuli. This also helped to quickly eliminate artificially generated images that could not produce more than one person in a photo, causing distortions to appear more clearly visible. Finally, as this study intended to measure adult facial perception, photos of children were excluded.

Once these strict exclusion criteria were defined, the large dataset was examined. The dataset was broken up into folders of 1,000 images per folder. Thus, images were culled through 1,000 photos at a time. When culling through images, the stimuli were coded for gender, age, smile strength, and race. Age was separated into two main categories: Young Adults (18-45) and Older Age Adult (46+). Race was indexed as "European" or "Non-European"

Smile strength was measured on a scale of 0-100 via an Instagram filter called "Smile Score.” This filter has been shown to reliably correlate with human ratings of smiles. These scores were then binned into three categories: “Low Smile” (Smile Score 0-33), “Moderate Smile” (34-66), and “High Smile” (67-100). Scores were binned to facilitate the creation of evenly matched stimuli sets between real and fake faces; some categorizations were later used for analysis. This process was repeated on each photo that fit the criteria until the desired demographically matched stimuli sets were assembled. Approximately 15,000 authentic images and 15,000 fake images were considered. The final approved image pool for the study was 362 images, all not violating the exclusion criteria. The finalized study survey stimuli set consisted of 48 images, randomly selected from the 362-image pool and evenly matched for age, gender, and affect (between the real and fake sets).

Results

Parametric Analysis of Artificial vs. Real Ratings We ran a 2 (subject gender) X 2 (artificial vs. real image) X 2 (gender of face image) X 2 (weak smile vs. strong smile) X 6 (images per block)[1] factorial ANOVA on the artificial versus real scale. Critically, there was no main effect of artificial vs. real facial ratings, F(1, 130) = 1.08, p = .30, η^2_p = .008. In other words, participants did not rate real or artificial faces as being

[1] Each block in this study consisted of 6 images within the same legitimacy*smile strength*gender group. For instance, there were six photographs of real people (legitimacy), with strong smiles (smile strength), who were male (gender). Therefore, these three variables were entered as random effects.

differentially more or less artificial (Artificial M (S.E.) = 4.65 (0.99), Real M = 4.58 (0.95)). This evidence can be taken to suggest that real and artificial faces are not differentiable.

This analysis did yield some significant effects and interactions. First, there was a significant main effect of smile intensity on ratings of artificiality, Mean smiling images = 4.8 (.097), non-smiling images 4.4 (.094), F(1, 130) = 55.85, $p < .001$, $\eta^2_p = .30$. Hence, smiling faces were rated as more "real" than faces expressing no smile. However, the critical analysis here is whether there was an interaction between image legitimacy (real or artificial) and whether the face was expressing a smile. This interaction was significant F(1,130) = 5.38, $p < .05$, $\eta^2_p = .04$, though small. People were slightly more likely to rate a smiling real face as real (Mean = 5.35 (0.067)) than a smiling artificial face not exhibiting a smile (Mean = 5.12 (0.056)).

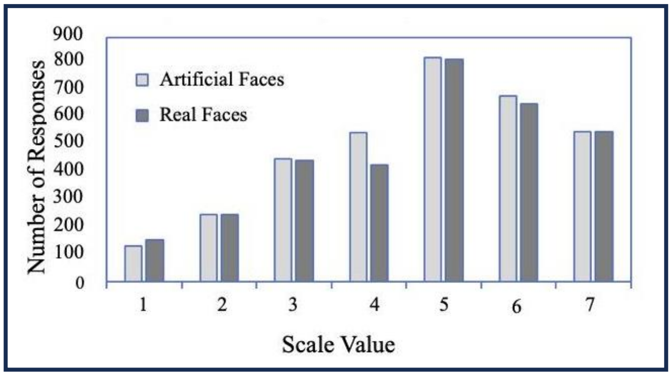

Figure 3: Histogram of responses in Experiment 1.

While we did not have specific predictions about gender, gender did play a significant role in the perceived legitimacy of faces. There was a significant main effect of facial gender F(1, 130) = 7.83, $p < .001$, $\eta^2_p = .057$. Male faces were rated as less real (Mean = 4.55 (.097)) than females (Mean = 4.68 (.091)). Hence, female faces were rated as more realistic than male faces.

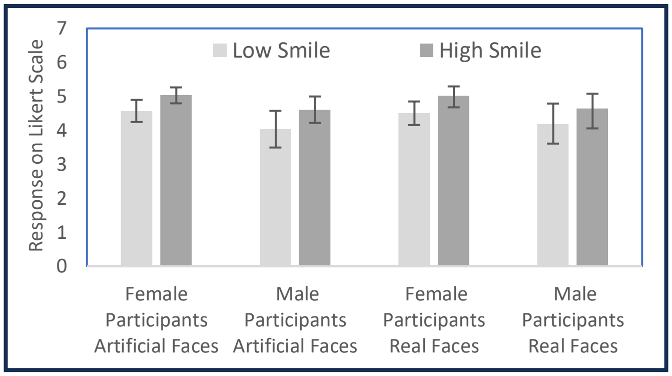

Figure 4: Mean responses by smile valence, participant gender, and real versus artificial.

Concerning between-participants main effects, there was a significant effect of participant gender F(1,130) = 5.2, $p =$ .024, $\eta^2_p = .039$, on real/fake rating. Although the effect is small, females rated faces as less realistic (M = 4.43, SD = 1.55) than men (M = 4.84, SD = 1.65). While these effects are interesting in themselves, we do not interpret them further because they were not part of the study's primary hypothesis. With respect to the age or race of the faces, all analyses failed to reach significance.

Non-Parametric Categorical Analysis To explore these data in greater detail, we conducted a follow-up categorical analysis examining whether participants were more likely to rate artificial faces as artificial than real. First, responses were coded as “artificial” if they were rated on the "artificial" side of the legitimacy Likert scale (a rating of 1, 2, or 3), as "real" if they were rated on the “real” side of the scale (a rating of 5, 6, or 7), or as “unsure” if they fell at the scale midpoint (4). We then categorized data further into bins based on whether, for each photo, each participant made a correct response (i.e., rating an artificial face as “artificial”), an incorrect response (i.e., rating an artificial face as a “real”), or an ambivalent response (“unsure”).

To explore these effects further, we analyzed data by examining the data non-parametrically. This is because the parametric analyses outlined above do not consider the accuracy of responses: did participants rate images of real and fake faces on the correct side of the legitimacy scale (i.e., if real faces rated as real, and fake as fake). As a first step, consider Figure 3, showing a histogram of participants' responses in Experiment 1. As can be seen in the graph, participants were far more likely overall to rate images as realistic (61% of all images) versus artificial (25%). This bias toward realism suggests that people generally believe these images are predominantly real, while in reality only half were real and half were artificial. This bias is highly significant (p < .000001, binomial test).

We performed an omnibus Chi-square test of independence to determine if condition (perceived legitimacy: correct, incorrect, or unsure) affected the frequency of responses $\chi^2(2) = 20.34$, $p < .001$, Cramer's V = 0.054. This suggests a significant difference between the three conditions in the frequency of correct or incorrect judgments (of artificial or real faces). However, this effect is primarily driven by the infrequency of “unsure” responses. Alternatively, only looking at binary choices between critical correct and incorrect judgments, a secondary Chi-square analysis omitting the Unsure category yielded no significant effect $\chi^2(1) = 2.52$, $p = .11$, Cramer's V = 0.02. This suggests that participants did not differ in accuracy depending on whether the faces were real or fake.

Discussion

Study 1 did not find strong evidence that college-age participants could readily discriminate real faces from artificially generated face stimuli. Yet there are limitations with the use of such convenience samples. For one, the participants in this group were mainly young, in their early twenties. Yet many of the stimuli in the experiment were of much older faces. It is possible that older participants with more life experience perceiving faces would show stronger effects.

Study 2 aimed to examine the effect of age congruity on discrimination between artificial and real images. A gap in prior literature exists for research examining identification accuracy between younger and older adults considering artificially generated images. However, in-group homogeneity provides a basis for the possibility that an age-based difference may exist regarding identification (e.g., Sporer, 2001). Research by Isaacowitz (2007) examined age differences in the recognition of lexical stimuli and facial expressions and found that older adults were less accurate in labeling certain emotions, relative to their younger counterparts. This indicates some differences in facial discrimination between age cohorts. Participant age was established as a variable and the effects of age-congruity vs. age-incongruity between participants and facial stimuli on discrimination accuracy was analyzed. We hypothesized that participants' accuracy in discriminating the authenticity of photos would be higher when a participant was within the same age bracket as the face pictured in the presented stimuli.

Study 2

Methodology Participants were pulled from Prolific, a paid subject pool. This study targeted those aged 18 - 40 (n= 76) and 50 and above (n=76). Participants were paid $5 to complete this 20-minute online anonymous Qualtrics survey. The stimuli and procedures used in Study 2 mirrored those in Study 1, except that once the survey was completed, participants were given a redemption code to enter Prolific to receive compensation.

Materials and Design: The materials and design used in Study 2 were identical to those used in Study 1.

Results

Parametric Analysis of Artificial vs. Real Ratings Unlike Experiment 1, Experiment 2 saw a significant main effect for whether an image was artificial or real, F(1,142) = 8.87, $p <$ .003, η^2_p = .059. However, the effect direction was opposite what was hypothesized: artificial faces were rated as more realistic (M = 4.82, SD = 1.22) than real faces (M = 4.64, SD = 1.32). The direction of this finding suggests that people were not able to accurately determine real from fake faces.

There was also a significant main effect for whether the stimulus face was young or old, F(1,142) = 40.58, $p < .001$, η^2_p = .22. The average rating of the young faces was 4.83 (between "Unsure" and "Probably Real"), versus 4.61 for the older faces (between the same response categories, closer to "Unsure"). This indicates that younger faces were perceived as less real than older faces. There was a significant interaction between whether a face was artificial or real and the age of the face, F(1,142) = 15.27, $p < .001$, η^2_p = .097.

There were three main effects: stimulus gender F(1.142) = 24.04, $p < .001$, η^2_p = .145 (Female M = 4.55 SE = 0.097 Male M = 4.68 SE = 0.091), facial age F(1,142) = 40.58, $p <$ .001, η^2_p = .222 (young faces M = 4.86 SE = 0.073 old faces M = 4.61, SE = 0.073), and participants age F(1,142) = 4.82, $p < .05$, η^2_p = .030 (young participants M = 4.55 SE = 0.087, older participants M = 4.64, SE = 0.082). We do not interpret these main effects as being relevant to the present analysis because their effects are so small and only reached significance due the large power inherent in our repeated measures design. There were several other higher-order interactions that were determined to be statistically significant, yet these were not effects that were part of the main hypotheses tested and their effect sizes were negligible.

There was a significant interaction between whether the image was artificial or real, whether the subject was young or old, and gender F(1.142) = 4.25, $p < .05$, η^2_p = .031. There was also a significant interaction between whether the face was artificial or real, facial gender, and whether the face was young or old F(1,142) - 30.635, $p < .001$, η^2_p = .177.

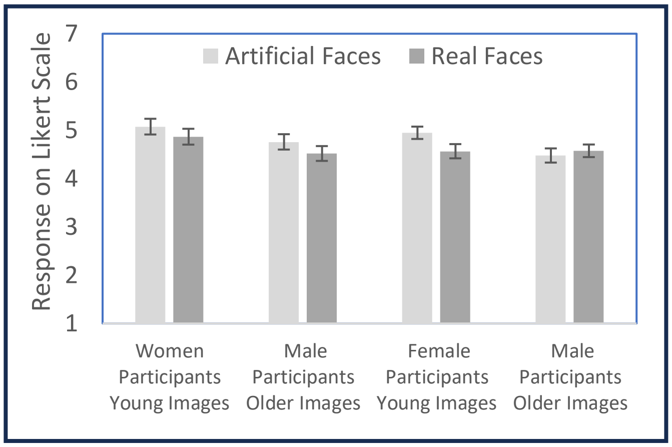

Figure 5: Mean responses by smile valence, participant gender, and real versus artificial.

Non-Parametric Categorical Analysis We conducted a non-parametric categorical analysis identical to that presented in Experiment 1. For the older sample, the result of the Chi-square test of independence using all three categories (correct, unsure, incorrect) was $\chi^2(2)$ = 658.77, $p < .0001$, Cramer's V = 0.425. Using just the two critical categories of correct versus incorrect, the result was nearly identical $\chi^2(1)$ = 656.11, p < .0001, Cramer's V = 0.462. For the younger sample, the effect was essentially the same: using all three categories $\chi^2(2)$ = 314.60, p < .0001, Cramer's V = 0.295, while using just correct and incorrect $\chi^2(1)$ = 310.13, p < .0001, Cramer's V = 0.318.

Discussion

The results of Experiment 2 suggest that both young and old participants were more likely to answer correctly (61% correct versus 20% incorrect, collapsed across young and old groups) when rating real faces versus artificial faces. When rating artificially generated faces, older and younger participants were both more likely to answer incorrectly (62% correct versus 22% incorrect). These findings suggest that, overall, the participants in Study 2 were more likely to respond that the faces they were viewing were real faces, regardless of whether the faces were truly real or artificial. The second study extends upon the findings in the first study by showing that participants of all ages cannot discriminate real from artificial faces.

General Discussion

Two experiments demonstrated that naïve participants are unable to discriminate photographs of real faces from AI generated faces. While there were some higher order interaction effects of age, gender, and emotional expression, none of these factors exhibited strong effect sizes and on the Likert scale we used the means were very close in magnitude. In sum, the significant finding here is that when images are controlled for the types of artifacts that provide obvious cues that an image is artificial such as background distortions and water marks, participants show no evidence of being able to detect whether a facial stimulus is of a real person or a product of generative AI.

We also examined several other factors. First, emotional expression. In Experiment 1 we found that the strength of the smile on stimuli faces affected ratings of how real or fake participants believed the images to be. Yet there was no significant effect of whether the images were real or artificial. In Experiment 2, along with the gender of the participant and that of the stimulus face, the facial stimulus' age and participant age affected participants' judgments as well. But none of these effect sizes were strong or followed a pattern that we had hypothesized.

In Experiment 1, where undergraduate students were sampled, there was no significant difference between ratings of real versus fake faces; however, in Experiment 2, where we used a more representative sample, participants did show a significant difference in their ratings of real and artificial faces. However, the observed effect was in the opposite direction as was hypothesized. Instead of rating the fake images as more fake, participants instead rated the fake images as more realistic than the real images in Experiment 2. The difference between expectations and findings may be a minor artifact, because the undergraduate sample was significantly younger than either of the samples in the second experiment, or due to different compensation methods (school credit versus lump sum). Nevertheless, in neither experiment did participant's ratings of how artificial or real images reached statistical significance, so we can interpret that if people can in fact discriminate real from fake faces, this ability was not demonstrated in responses to our study. One explanation for this finding is simply that the computer program that produced these faces created such realistic faces that there are no visible cues to the veracity of the image. In a debriefing conversation with participants in the Rutgers sample, many admitted a complete inability to discern whether the images were of real people (to the point that many participants were dismayed, some claiming that they thought all the faces were real).

There are several implications of these findings. First, AI programs have become so sophisticated and commonplace that they open a Pandora's box of problems for both individuals and society. While the ability to mask identity through costume, makeup, and photographic manipulations have existed now for over a century, programs such as StyleGAN present a new challenge as the program rapidly creates novel images of faces unrelated to any living individual. This technology could be easily used by individuals with nefarious intentions, such as in creating fake identities in the application of crime or fraud. Indeed, already there have been cases of fake journalists in the Ukraine (Twomey, 2023) and Israel conflicts (Kabbaje, 2023) where artificially generated images of individuals were used in press reports, and yet under scrutiny established that the individuals never had existed at all and were likely part of a disinformation campaign. While in the past it would be possible to use random pictures downloaded from the internet to fabricate a new identity, with reverse image search such frauds were easily established. Yet not so with StyleGAN. The code to run the program is a simple freeware python application which can readily run on a computer with a fast graphics card. Once installed, the program can be input with a novel set of facial images and start producing a new set of AI generated stimuli. The ease with which it produces images of people who have never existed on earth opens the door to fraud, misrepresentation, and manipulation.

Because of this, people need to be better educated that this technology exists and how to look for cues within images themselves for evidence of their artificial origin. Recall that Bergstrom & West (2022) found that people became more accurate over time when discriminating real from fake faces in their online experiment. Even though most people peaked around 60-70% accurate discriminations, there was evidence within the data of learning cues that could be used to establish whether the images were real or not. However, because their stimulus set included images with watermarks and distortions, it may simply be that their participants were informed by these features. In the present study, we took care to only use images where those features were absent, yet we did not provide accuracy feedback which precluded understanding whether participants were learning. Providing such accuracy feedback may be an important consideration for future research.

A second issue is that while the present study used facial stimuli, the AI techniques used here can and are being extended to other types of non-social stimuli. (Hanna, 2023; King, 2022).

Finally, the technology has advanced so quickly since Fall 2023 that there are now programs that allow not only still photo output but video and audio stimuli as well. While the present study demonstrates that this technology has matured and improved, more research is necessary to determine how to train individuals to better recognize the products of generative AI.

References

Ackerman, J. M., Shapiro, J. R., Neuberg, S. L., Kenrick, D. T., Becker, D. V., Griskevicius, V., ... & Schaller, M. (2006). They all look the same to me (unless they're angry) from out-group homogeneity to out-group heterogeneity. *Psychological science*, *17*(10), 836-840.

Arias, P., Soladie, C., Bouafif, O., Roebel, A., Seguier, R., &amp; Aucouturier, J. J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. IEEE Transactions on Affective Computing, 11(3), 507-518.

Bergstrom, C., & West, J. (2022). Discussion of data related to the "Calling Bullshit" website. Personal communication.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception?. *Psychological review*, *105*(3), 482.

Girard, J. M., Shandar, G., Liu, Z., Cohn, J. F., Yin, L., &amp; Morency, L.-P. (2019). Reconsidering the Duchenne smile: Indicator of positive emotion or artifact of smile intensity?. 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 594-599.

Gunnery, S. D., & Ruben, M. A. (2016). Perceptions of Duchenne and non-Duchenne smiles: A meta-analysis. *Cognition and Emotion*, *30*(3), 501-515.

Hanna, D. M. (2023). The Use of Artificial Intelligence Art Generator "Midjourney" in Artistic and Advertising Creativity. *Journal of Design Sciences and Applied Arts*, *4*(2), 42-58.

Hsu, C. C., Lee, C. Y., & Zhuang, Y. X. (2018, December). Learning to detect fake face images in the wild. In *2018 international symposium on computer, consumer and control (IS3C)*. IEEE.

Isaacowitz, D. M., Löckenhoff, C. E., Lane, R. D., Wright, R., Sechrest, L., Riedel, R., &amp; Costa, P. T. (2007). Age differences in recognition of emotion in lexical stimuli and facial expressions. *Psychology and Aging, 22(1),* 147–159.

Israel, M. J., & Amer, A. (2023). Rethinking data infrastructure and its ethical implications in the face of automated digital content generation. *AI and Ethics*, *3*(2), 427-439.

Kabbaj, S. (2023). A Call for Responsibility: Social Media and the Need to Monitor Hate Speech and Fake News for Global Peace. *Journal of Information Systems and Technology Research*, *2*(3), 115-123.

Karpathy, A., Abbeel, P., Brockman, G., Chen, P., Cheung, V. Duan, Y., Goodfellow, K., Kingma, D., Ho, J., Rein, H, Salimans,T. ,Schulman, G., Sutskever, I., Zaremba, W. (2016). *Generative Models.* OpenAI White Paper.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. pp. 4396–4405.

Karras, T., Aittala, M., Laine, S., Harkonen, E., Hellsten, J, Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Proceedings of the Advances in Neural Information Processing Systems.*

King, M. (2022). Harmful biases in artificial intelligence. *The Lancet Psychiatry*, *9*(11), e48.

Levesque, H. J. (2017). *Common sense, the Turing test, and the quest for real AI*. MIT Press.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89-94.

Mondloch, C. J., Lewis, T. L., Budreau, D. R., Maurer, D., Dannemiller, J. L., Stephens, B. R., & Kleiner-Gathercoal, K. A. (1999). Face perception during early infancy. *Psychological science*, *10*(5), 419-422.

Montagne, B., Kessels, R. P., Frigerio, E., de Haan, E. H., & Perrett, D. I. (2005). Sex differences in the perception of affective facial expressions: do men really lack emotional sensitivity?. *Cognitive processing*, *6*(2), 136-141.

Neufeld, E., & Finnestad, S. (2020). In defense of the Turing test. *AI & Society*, *35*, 819-827.

Olderbak, S., Wilhelm, O., Hildebrandt, A., & Quoidbach, J. (2019). Sex differences in facial emotion perception ability across the lifespan. *Cognition and Emotion*, *33*(3), 579-588.

Scott, L. S., & Monesson, A. (2009). The origin of biases in face perception. *Psychological Science*, *20*(6), 676-680.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law*, *7*(1), 36.

Tariq, S., Lee, S., Kim, H., Shin, Y., & Woo, S. S. (2018, January). Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd international workshop on multimedia privacy and security* (pp. 81-87).

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148.

Turing, A. (1950). Computing Machinery and Intelligence, *Mind*, **59,** 433–460.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos one*, *18*(10), e0291668.