# How do the Socio-Cognitive Impacts of Real vs. AI-Generated Facial Expressions of Emotion Differ?

**Megan Lawrence**

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Ph.D. in Experimental Psychology

School of Psychology
Faculty of Social Sciences
University of Ottawa

## ABSTRACT

Media generated by aritificial intelligence (AI) is becoming ubiquitous. As it becomes more realistic and more common-place, it is important to know whether we can distinguish it from real media, how accurately it replicates real media, and what negative socio-emotional effects it could be perpetuating. A critical concern regarding AI-generated faces is that the datasets used to train the generators consist largely of images of White men. When the training sets are biased, the output will be biased. Questions then arise regarding whether AI generators are able to generate faces of varying genders and races with equal verisimilitude. Additionally, do they generate different facial expressions accurately? The three studies presented in this thesis examined these questions with regards to one AI-generator, which uses an implementation of Stable Diffusion. The first study examined how well people can distinguish AI-generated faces from real faces across genders (men and women), races (Asian, Black, and White), and emotional expressions (anger, fear, happiness, and sadness). Examination of d-prime values revealed that, although detection of AI-generated faces was high overall, participants were particularly good at identifying AI-generated faces of people of colour and men. Criterion was also examined and revealed that participants had a bias towards classifying all faces except for those of White men as AI-generated. The second study examined differences in emotion detection for AI-generated faces vs. real face photographs. Participants viewed faces of varying genders (men and women), races (Asian, Black, and White), and emotional expressions (anger, fear, happiness, and sadness) and rated their emotions using scales for the six basic emotions plus neutral (neutral was denoted by setting all emotion scales to zero). These ratings were then transformed into intensity (the degree of the target emotion rating), saliency (the ratio of the target emotion rating to all ratings), and accuracy (whether the target emotion was the highest rated emotion) scores. Results indicated that, broadly speaking, AI-generated images were lower than their real face photograph counterparts on all dependent variables. However, there were some subtleties. Specifically, the AI-generator seemed to perpetuate the stereotype of "the angry Black woman", as AI-generated images of Black women expressing anger were more intense, salient, and accurately identified than their real counterparts, and this was not seen for other gender/race conditions expressing anger. Finally, the third study examined whether social attributions assigned to AI-generated faces would differ from those assigned to real face photographs using a mixed measures approach. Participants viewed AI-generated and real face photographs of varying genders (men and women), races (Asian, Black, and White), and emotional expressions (anger and happiness) and gave open ended first impressions. Responses were then coded quantitatively on four dependent variables (emotional valence, emotional arousal, warmth, and dominance). Results again indicated that the AI-generator seemingly perpetuated commonly held stereotypes, especially with regard to gender and race. Again, the stereotype of the "angry Black woman" was prominent. Additionally, the stereotype of the "passive Asian woman (and sometimes man)" was also apparent. Taken together, these studies provide evidence that AI-generated media, and in turn the training sets used to create these generators, are indeed biased against generating accurate depictions of people of colour, particularly women of colour.

## ACKNOWLEDGEMENTS

First and foremost, I have to thank my wonderful supervisor, Professor Charles Collin, who has been nothing but supportive at every point throughout my degree. Not only has he been an amazing mentor in research, but he also showed so much compassion throughout the strange path I took getting here; understanding when things were too much and I needed to take a step back, and welcoming me back into the lab when I was ready with no hesitation. For his support I am so grateful and with his guidance I am truly proud of the work I was able to accomplish in my time in his lab.

I would also like to give a big thank you to my amazing committee members, Professors Denis Cousineau, Patrick Davidson, and Jean-Philippe Thivierge. Their encouragement and feedback throughout this process made my work better, and made the way I approach research better.

I am also deeply grateful to all of my labmates (current and past) who sat through countless iterations of my studies, provided feedback, and were there to just chat when I needed a break. None of my work would have gotten done if they hadn't tested and re-tested everything, so thank you from the bottom of my heart. And a special thank you to Kasey and Emily, I am so grateful to have worked with you both.

From my personal life, I would like to thank my amazing husband, Joven, who has supported me unconditionally throughout my degree. From offering words of encouragement when things were hard, to listening to me ramble for hours about how cool I think my work is, his belief in me never wavered. For that I will forever be grateful. Again, my thesis would not be what it is without his support. Thank you to my family who also never stopped believing in me and always listened to me ramble about my studies, even when I was making no sense. Finally,

thank you to all of my friends who have supported me. Jenna, Ruth, and Taryn, not only for encouraging my work, but also for reminding me of the importance of taking time off and spending it with the people I love. Adelaide, Vanessa, Alex, and Marilou, from coffee dates to writing sessions, without them I don't think I would have left my house until my thesis was fully written. They kept me sane throughout this process, and I am so thankful to have gotten to experience this program with so many amazing people, and the friendships that have come out of it are some I will cherish forever.

# TABLE OF CONTENTS

# 1. GENERAL INTRODUCTION

Applications for automatically creating media via Artificial Intelligence (AI) are advancing at a rapid pace. Generative Adversarial Networks (GANs) and deepfake applications are rapidly evolving and becoming more accessible to the public (Goodfellow et al., 2020; K. Wang et al., 2017; Whittaker et al., 2020). With software like this, altered and entirely fictional media (audio, image, and video) can be created rapidly and with ease. And, while these types of applications can be used for positive purposes, such as enhancing existing data, removing language barriers (Whittaker et al., 2020), and aging the faces of missing children to help find them (Chandaliya & Nain, 2022), many negative uses also arise. These include an exacerbation of the spread of misinformation, an increase in the creation of non-consensual intimate imagery (i.e., revenge porn), an undermining of trust in genuine recordings of poor behaviour, and much more (FBI, 2023; Whittaker et al., 2020). It is thus important to be able to detect such images.

While software tools are being developed to do this, the most accessible and common way to detect AI-generated media is through human perception. Not every person will be able to use detection software, but everyone can try to detect AI-generated media using their own knowledge and perceptual abilities. However, the human ability to detect AI-generated media, specifically face images, may not be very good. Indeed, numerous studies have shown that the ability to detect AI-generated images is at about chance-level (Bray et al., 2023; Josephs et al., 2023; Moshel et al., 2022; Nightingale & Farid, 2022; Rossi et al., 2023; Tucciarelli et al., 2022; X. Wang et al., 2022). However, one study by Rossi et al. (2022) found that accuracy in detecting AI-generated twitter accounts was as low as 10% chance, whereas detecting the real accounts was drastically higher (58.5-91.4%). This underscores the importance of not only

looking at average detection rate, but specifically the ability to distinguish between real and AI-generated media.

The work described below examines issues related to this rapidly advancing technology. In particular, I examine how discriminable facial expression images generated by AI are from real images, and how they affect the emotional judgements and social cognitions of people viewing them. This was be done by showing participants real and AI-generated images, and having them attempt to distinguish between the two, rate the emotional expression intensities of the images, and give first impressions of the people in the images. In some studies, the information the participants were given about the presence of AI images was varied. Below, I first briefly discuss some relevant theoretical models and previous literature on these topics before describing in detail the three studies I executed for my thesis work.

### 1.1. Theoretical Models and Their Applications

The Emotions As Social Information (EASI) model states that we make inferences from and react to others' facial expressions, and that these inferences and reactions guide our behaviours toward them (Van Kleef, 2009). For example, one might see an individual frowning and infer that they are feeling sad. Their expression may thus elicit compassion within the observer, which, in turn, cultivates supportive behaviour. This model begins with the decoding of emotions. However, the EASI model does not specify how emotions are decoded. We must therefore look to models of emotion decoding in combination with the EASI model to gain a more complete understanding.

One highly influential model in this area is Ekman's (1989, 1992) model of the six basic emotions. While it has been argued that emotions are culturally specific, some evidence also suggests that there are universals. Specifically, Ekman (1992) argues for the existence of six

basic emotions: anger, disgust, fear, happiness, sadness, and surprise. This model argues that these emotions are both universal and discrete. Further, Ekman suggests that these expressions each involve a unique sets of facial muscle activations, called Action Units. For instance, a genuine smile involves flexion of the orbicularis oculi and zygomaticus major. It is via these action units that expressions are transmitted and decoded.

Several other models guided this research, particularly Study 3, in terms of qualitative data coding. The models selected all use two-dimensional structures, which place the variables of interest on separate, bipolar axes. To code valence and arousal, a two-dimensional semantic structure of affect model was applied (Feldman Barrett & Russell, 1998). Similarly, when considering warmth and dominance ratings, the Revised Interpersonal Adjectives Scale (IAS-R; Gurtman & Pincus, 2000) was used, as well as the Interpersonal Circumplex model (IPIP-IPC; Markey & Markey, 2009). Feldman, Barrett, and Russel's (1998) model of valence and arousal (or in their words, affect and activation) posits that valence and arousal are independent of one another: positive and negative for valence, vs. activated and deactivated for arousal. The IAS and IPIP-IPC models both provide a similar structure for warmth and dominance as the valence-arousal model. Warmth and dominance are placed on separate, bipolar axes, and create intermediate categories between each axis. Both models follow the same broad structure, but where the IAS, as its name would suggest, uses adjectives and places them in the circular model, the IPIP-IPC presents an alternative that was designed to be more easily understood by those using the scale (Markey & Markey, 2009).

Using the above models as guides, more refined coding schemes were developed to examine the broad categories of emotional valence, arousal, warmth, and dominance. These models were selected after viewing the data (see Section 4.3.4. Coding Scheme for more details).

We can combine the EASI model, Ekman's model of emotions, and our measures of emotional valence, arousal, warmth, and dominance when examining faces. Faces are a rich source of social and emotional information (e.g., identity, emotional state, attractiveness, attention, etc.) from which we make inferences about a person's personality traits (Todorov et al., 2015; Van Kleef, 2009). Research on the social attributions we assign to an individual based on viewing their faces is quite vast. There are several factors that impact social perceptions of an individual and the attributes we assign them. These include facial expression, age, attractiveness, race of the individual in the photo, race of the viewer, etc., some of which create an "other group" effect (Campbell et al., 2010; Cortes et al., 2019; Kiiski et al., 2016; Todorov et al., 2015; Yan et al., 2016; Young et al., 2011). Despite the complexity introduced by these many factors, or perhaps because of them, there is a large amount of research on the social and emotional effects of faces. For example, some authors have explored the emotional enhancement of memory effect, whereby emotionally expressive faces are remembered better than neutral faces (Anderson et al., 2006; Baudouin et al., 2000; Bradley et al., 1997; Carstensen & DeLiema, 2018; Charles et al., 2003; Chen et al., 2015; Collin et al., 2022; Grady et al., 2007; Jackson et al., 2008; Liu et al., 2014; Tay & Yang, 2017). This effect may interact in complex and subtle ways with other factors, such as the age of the viewer. Some research finds that young adults show a negativity bias in memory, such that they remember angry and fearful faces better (Anderson et al., 2006; Chen et al., 2015; Grady et al., 2007; Jackson et al., 2008) whereas older adults show a positivity bias, such that they remember smiling faces best (Carstensen & DeLiema, 2018; Charles et al., 2003).

### 1.2. What is AI-generated media and how is it created?

AI-generated media includes images, songs, text, and a variety of other creations that are made, in whole or in part, via applications that take advantage of deep learning algorithms trained on large datasets of media. Among the most popular types of applications used for this purpose are deepfake generators, GANs, and Stable Disffusion models. Originally, deepfake images were created by hobbyists, and they were quite obviously doctored. These hobbyists would, for instance, insert the faces of female celebrities onto the bodies of actors in adult videos. As this gained popularity, and the code was shared publicly online, deepfakes became more advanced, and more difficult to detect. Now, the tools exist for easily and quickly creating fake media across many formats, including audio, images, videos, or a combination of them all (Whittaker et al., 2020).

GANs are deep learning models that are created using an iterative process, by which they get better at creating artificial media. This is done by using two linked systems: a generator and a discriminator. The system is based on game theory. The generator is given a training set of data, which it uses to alter or create new media. The discriminator is then given real media and generated media and must distinguish between them. Each time the discriminator correctly classifies a piece of generated media, the generator's weights are updated, and it improves. If the discriminator is fooled, and incorrectly classifies a piece of generated media, it learns, so that it does not make that mistake again. Through this process, both the generator and discriminator improve. Once trained, the generator will be retained and can now be used to create new media. Usually, the discriminator will be discarded at this point (Goodfellow et al., 2020; K. Wang et al., 2017; Whittaker et al., 2020). Diffusion models are rapidly becoming ubiquitous and generally produce images that are more realistic than those created with GANs. They function by

introducing noise into their training sets, and gradually removing that noise to generate new media (Dhariwal & Nichol, 2021).

The result of this process is a computer program that can take a text prompt and alter or create entirely new media. For example, the software I used in my studies, Leonardo.Ai, has both functions. The first is to create entirely new images based solely on the text prompt. Users type in key words that describe the image to be generated, as well as additional settings, such as the style of image wanted (i.e., photography, illustration, etc.), and the generator will output several options of images. The other function Leonardo.Ai has, which I used (see section 2.3.1. Stimulus Creation for full explanation), is image-to-image generation. Users can upload their own images to use in combination with key words and additional settings to guide the generation of a new image or alter the existing image.

### 1.3. Overview of previous AI literature

Previous work has examined the human ability to accurately distinguish between real and AI-generated faces, and has found that it has been declining as generative AI technology has improved (Bray et al., 2023; Duffy et al., 2024; Josephs et al., 2023; Lago et al., 2022; Miller et al., 2023; Moshel et al., 2022; Nightingale & Farid, 2022; Rossi et al., 2023; Tucciarelli et al., 2022). AI-generated media, more broadly, now have fewer artifacts and are overall more realistic than before, with some authors finding hyperrealism of AI-generated faces (Miller et al., 2023). However, this is not equally true for all kinds of faces. For example, Nightingale and Farid (2022) found that the AI-generated faces of White men, and women to a slightly lesser degree, were particularly difficult to distinguish from real photographs, while face images of people of colour were more accurately classified. They hypothesize that this is because the training sets for AI-generators are largely biased towards White, adult men.

Most research to date examining distinguishability, emotion recognition, and social perceptions of AI-generated faces has focused solely on images of White individuals (Eiserbeck et al., 2023; Moshel et al., 2022; Rossi et al., 2023; Tucciarelli et al., 2022). Within this research, much of it still finds about chance-level performance when classifying real vs. AI-generated faces (Moshel et al., 2022; Rossi et al., 2023; Tucciarelli et al., 2022). Where studies do include a variety of races, for example, Bray et al., 2023, they often do not include the factor of face race in their analyses. Rossi et al. (2022) explicitly did not include a variety of races so as to provide more control over their experiment. While this does reduce the number of extraneous factors impacting the results, ecological validity can be improved by including photos of individuals of different races, and this can be a factor in analyses.

More recent work has highlighted the importance of examining a variety of factors, such as gender and race, and how AI-generators perpetuate and enhance stereotypes (AlDahoul et al., 2025; Assis & Moura, 2025; Bhardwaj et al., 2025; Bianchi et al., 2023; Ghosh & Caliskan, 2023; Jääskeläinen et al., 2025; Lawrence & Collin, 2025; Lawrence et al., 2025; Locke & Hodgdon, 2025; Nightingale & Farid, 2022; Wu et al., 2025). One article examining the datasets that Stable Diffusion is trained on found that the representation of different races is skewed, and that it gets skewed even further when implemented by Stable Diffusion (AlDahoul et al., 2025). As such, the authors created their own generator and found that theirs, which was trained using the FairFaces dataset (Kärkkäinen & Joo, 2021), generated images of varying races with much more equal frequency. This work, however, did not examine the realism of their generated faces. Moreover, when examining generators, research finds that, not only are women and people of colour (POC) under-represented when prompts including “person” are used (Assis & Moura, 2025; Ghosh & Caliskan, 2023; Jääskeläinen et al., 2025; Wu et al., 2025), but the images

generated are often more sexualized (Park, 2024). One article found that the prompt "person" yields mostly light-skinned men, while prompts for women with the qualifiers "American" or "European" yielded light-skinned headshots of women, and prompts for women from countries with predominantly POC populations yielded full-body images with accentuated breasts and hips, perpetuating the sexualization of women of colour (Ghosh & Caliskan, 2023). Moreover, AI-generators perpetuate societally held biases, stereotypes, and norms. For example, when generating images of different professions we see gender roles and racial stereotypes reinforced (Gorska & Jemielniak, 2023; Locke & Hodgdon, 2025; Mubashir, 2024), and often the proportions do not align with real world ratios of people in those professions (Chauhan et al., 2024; Gawali et al., 2024). In addition, when generating images depicting wealth vs. poverty, yet again, racial stereotype emerge, where wealth is often represented by White men in the Global North, and poverty depicted as POC and the Global South (Dehouche, 2021; Jääskeläinen et al., 2025).

An individual's knowledge or presumption that an image is AI-generated vs real may impact certain evaluations of the face, however there is conflicting evidence. In one study, when faces were *presumed fake* (even though all images were real), there were no differences in how they were perceived for the angry facial expression. For happiness, however, the presumed fake faces did not follow the pattern expected and demonstrated for real faces. In fact, presumed fake happy faces did not differ from neutral faces, and did not elicit differences in the P1, N170, and EPN components (Eiserbeck et al., 2023). Another article found that labelling faces as either AI-generated or made by an artist did not impact responses to questions examining racial biases (AlDahoul et al., 2025). How knowledge of whether an image is AI-generated or not impacts perceptions of that image is unclear, and more research is needed to make firm conclusions.

### 1.4. Overview of the current thesis

The following studies extend previous research in several ways. First, they examine the human ability to distinguish between real and AI-generated faces expressing a wider variety of emotions. Most research to date has used almost exclusively images of individuals smiling or with a neutral expression (Bray et al., 2023; Gragnaniello et al., 2022; Lago et al., 2022; Liefooghe et al., 2023; Moshel et al., 2022; Nightingale & Farid, 2022; Rossi et al., 2023; Shen et al., 2021; Tucciarelli et al., 2022; X. Wang et al., 2022). Even when there is slight variation in facial expressions, it has not been experimentally manipulated or analysed. This is, perhaps, because many AI generators are trained using datasets that primarily contain smiling individuals. For this reason, for Study 1, we used a generator that had been weighted towards images balanced on emotional expression, as well as other variables of interest such as race and gender. For studies 2 and 3, a newer model had been released which did not support refining models, and therefore we did not continue to use this model.

Second, the studies outlined below included a wider range of races among the stimulus images. To reiterate, although more work has been published since data collection of the studies outline below (AlDahoul et al., 2025; Assis & Moura, 2025; Bhardwaj et al., 2025; Bianchi et al., 2023; Ghosh & Caliskan, 2023; Jääskeläinen et al., 2025; Locke & Hodgdon, 2025; Nightingale & Farid, 2022; Wu et al., 2025), at the time of our data collection, much of the research to date had used stimuli mostly representative of White men (Moshel et al., 2022; Rossi et al., 2023; Tucciarelli et al., 2022).

Lastly, AI-generated images of women's faces, even among White subjects, were more easily discriminated from photographs (Nightingale & Farid, 2022). Therefore, as well as the previously described weighting of the generator, it was weighted so that women's and men's

faces were balanced. We included equal numbers of men and women of the three races we included in our stimulus set (White, Black and Asian), expressing the various emotions. To reiterate, for studies 2 and 3, using this weighted generator was not possible due to the new model not supporting this function.

Refining the generator was done by using faces drawn from the Racially Diverse Affective Expressions (RADIATE; Conley et al., 2018; Tottenham et al., 2009), a database that includes men and women who self-identified as Asian, Black, Hispanic, and White. Each individual posed eight facial expressions (anger, calm, disgust, fear, happiness, neutrality, sadness, and surprise), each with opened and closed mouth variations. That is, the facial expressions in the RADIATE database were not natural expressions, rather the individuals were instructed on how to pose each expression and given time to practice before being photographed. For our work, we used the four open-mouthed variations of emotional expressions of anger, fear, happiness, and sadness. See section 2.3.1. Stimulus Creation for details on how the generator was refined.

In summary, the three studies in this thesis will expand this area of research by examining several facets of human perception of AI-generated faces across different genders, races, and emotional expressions.

Lawrence, M., Cimermanis, K. N. E. & Collin, C. A. (2025). Not all AI-generated faces are created equal: impacts of model gender, race, and emotional expression on classification accuracy. *AI & Soc*. https://doi.org/10.1007/s00146-025-02670-7

## 2. STUDY 1: Not all AI-generated faces are created equal: impacts of model gender, race, and emotional expression on classification accuracy

Megan Lawrence*, Kasey N. E. Cimermanis, & Charles A. Collin

### 2.1. Abstract

Generative Artificial intelligence (AI) technologies allow users to create novel images and modify existing images easily and rapidly. However, the training sets used to create the generative technologies may be biased; they may be primarily trained on images of White men (Kärkkäinen & Joo, 2021, https://doi.org/10.1109/WACV48630.2021.00159). When the training is biased, the output will also be biased. In this case, generating realistic faces of White men may be done with ease, but varying race, emotional expression, or gender may reduce realism. To determine the degree to which this bias exists, AI-generated faces were created using Leonardo.ai's image-to-image algorithm, with two genders (man and woman), three races (Asian, Black, and White), and four emotional expressions (anger, fear, happiness, and sadness). N = 138 participants were presented with real and AI-generated faces and classified them as such. A 2 × 3 × 4 repeated measures ANOVA was run with sensitivity (d') as the dependent variable. Participants were significantly better at identifying AI-generated faces of people of colour, more specifically, men of colour. Criterion values were also examined; participants had a bias towards classifying all faces except for those of White men as AI-generated, which they tended to classify as real. Our results support previous findings showing that demographic biases exist in generative AI tools and expand on those by showing how such biases interact with the emotional expression of the face.

## 2.2. Introduction

AI-generated media is increasingly easy to create, with Generative Adversarial Networks (GANs), diffusion models, and deepfake applications rapidly evolving and becoming widely available to the public (Goodfellow et al., 2020; K. Wang et al., 2017; Whittaker et al., 2020). Diffusion models are rapidly becoming ubiquitous and generally produce images that are more realistic than those created with GANs. They function by introducing noise into their training sets, and gradually removing that noise to generate new media (Dhariwal & Nichol, 2021).With this ease of access and rapid advancements in the technology, there is a concomitant increase in both the negative and positive uses of it (Chandaliya & Nain, 2022; FBI, 2023; Whittaker et al., 2020). For this reason, understanding our ability to distinguish between real and fake media becomes of paramount importance.

Although there is software that can detect AI-generated media, the most accessible and direct way to do so is through human perception. While a few studies have examined whether we are able to distinguish between real and AI-generated faces, the research is still limited, and AI generation is advancing rapidly. Most research to date finds that our accuracy at classifying real vs. AI-generated faces is at about chance level (Bray et al., 2023; Josephs et al., 2023; Moshel et al., 2022; Nightingale & Farid, 2022; Rossi et al., 2023; Tucciarelli et al., 2022). However, one study by Rossi et al. (2023) found that while average accuracy scores are at about chance level, when examined separately, accuracy scores for labelling AI-generated media are far below chance level, while accuracy scores for labelling real media range from about chance level to far above chance. In their study, participants were shown a mix of real and AI-generated Twitter accounts as if they were scrolling through a Twitter feed. Participants were asked to label the profiles as real or bots (i.e., AI-generated), rate the likelihood of the accounts being bots, and rate

different components that caused them to be suspicious of an account being a bot. Accuracy in labelling bots ranged from 10-27.4%, accuracy in labelling real accounts ranged from 58.5-91.4%, and overall accuracy was 48.9%.

While the studies outlined above all found similar results, they are not very generalizable. This is because research in this area has generally not been very inclusive. The base models used for AI-generation are created using biased image sets. The images used are often arbitrarily scraped off the internet, as they are easily accessible. Because of this, image sets are often biased towards generating realistic images of White faces as they are more prevalent in the training sets, while other races[1] are underrepresented (Kärkkäinen & Joo, 2021). For example, many past studies have not included people of colour in their stimuli (Moshel et al., 2022; Rossi et al., 2023; Tucciarelli et al., 2022), or the AI generator used has not been trained on an inclusive set of images, and thus the stimuli showing people of colour are not comparable to those of White people (Nightingale & Farid, 2022). In addition, some studies that do include people of colour simply do not report results based on race (Bray et al., 2023). The present study aimed to address this issue by using images of Asian, Black, and White models. By doing so, we were able to examine differences in the quality of image generation for Asian and Black models compared to White models. Finally, research in this area has not examined the efficacy of AI-generation of different facial expressions. The present study aimed to address this issue by examining participants' ability to distinguish real from AI-generated faces portraying a range of different facial expressions.

---

[1] The term race is used throughout based on APA guidelines defining race as "social construction and categorization of people based on perceived shared physical traits that result in the maintenance of a sociopolitical hierarchy" (American Psychological Association, 2019, p. 47).

Since stimulus creation and data collection, more work has been published which provides more evidence for the above statements. For example, several recent articles examined whether using generic prompts resulted in biased and stereotyped images (AlDahoul et al., 2025; Assis & Moura, 2025; Bianchi et al., 2023 Chauhan et al., 2024; Gawali et al., 2024; Ghosh & Caliskan, 2023; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Locke & Hogdon, 2024; Mubashir, 2024; Park, 2024; Wu et al., 2025), and, indeed, they found stereotyped portrayals of people of colour and women.

For all of our stimuli we created refined models using the Leonardo.ai site's tools for doing so. This allowed the creation of a customized generative AI model based on a user-provided sample of stimuli. This was done to increase the realism of all the generated images. A total of 24 refined models were created, one for each of the 3 races (Asian, Black, White) × 2 genders (male, female) × 4 expressions (Happy, Sad, Angry, Fearful) that were examined in this study. More details are given in the methodology section.

As this is a new and rapidly evolving field, it was difficult to make firm predictions. However, past research has indicated that humans can only detect AI-generated faces at about chance level (Bray et al., 2023; Josephs et al., 2023; Moshel et al., 2022; Nightingale & Farid, 2022; Rossi et al., 2023; Tucciarelli et al., 2022). The following research hypotheses (H) guided exploration of the topic. If our refinement of the AI generator works as intended, this prediction should generalize across races and genders. However, it is possible that the refinement is still not enough to balance the initial training of the generator. With this in mind, the following two hypotheses were formed:

**H1:** Because AI generators are trained disproportionately on White face stimuli (Kärkkäinen & Joo, 2021), they will be biased towards generating more realistic White faces

compared to faces of people of colour. We therefore predict that AI Faces (AIFs) of people of colour will be more accurately distinguished from real faces than will AIFs of White people.

**H2:** Because AI generators are trained disproportionately on images of men's faces (Kärkkäinen & Joo, 2021), they will be biased towards generating more realistic faces of men compared to faces of women. We therefore predict that AIFs of women will be more accurately distinguished from real faces than will AIFs of men.

Although the generator was refined for facial expression as well, there is very little research which examines the effect of facial expression of AIFs. As this is the case, the following research questions (RQ) were generated to help guide further exploration of the behavioural data:

**RQ1:** Will AIFs with certain facial expressions be less accurately distinguished than AIFs with other facial expressions? It is possible that the stimulus training sets of AI generators contain disproportionately more of certain expressions than others (e.g., more smiling faces as compared to angry, fearful, and sad ones). This would potentially cause them to generate more realistic images of certain expressions compared to others.

**RQ2:** Will the factors of race, gender, and expression interact in determining the verisimilitude of AIFs? The lack of research in this area makes it difficult to form firm hypotheses about interactions of the above effects. However, we expect that they may interact with each other such that, for example, an AIF of a sad woman of colour will be more accurately distinguished than an AIF of a happy White man. However, it is not clear to what degree the different factors (race, gender, expression) would interact rather than producing additive effects.

## 2.3. Methods

### *2.3.1. Stimulus Creation*

The first step in this experiment was to create the stimuli. The Real Faces (RFs) were drawn from the RADIATE database (Conley et al., 2018; Tottenham et al., 2009). It contains images of men and women who self-identify as Asian, Black, Hispanic, or White. The individuals express eight emotions (anger, calm, disgust, fear, happiness, neutrality, sadness, and surprise) with closed-mouth and open-mouth variations. A subset of these RFs were used to generate the AIFs using an AI algorithm known as image-to-image generation, which is part of Leonardo.Ai (*Home | Leonardo.Ai*, n.d.). Leonardo.Ai is an online tool that implements diffusion models to allow easy image generation. In this case we used its implementation of Stable Diffusion 2.1 (*Home | Leonardo.Ai*, n.d.), as informal testing showed that this yielded the most realistic face images. The RFs were submitted to the program and using keywords specific to the condition, for example, "a happy black woman", AIFs were generated with four facial expressions: angry, fearful, happy, and sad (all open-mouthed variants). The stimulus set contained images portraying men and women of three different races (Asian, Black, and White). See Figure 1 for example stimuli. This figure shows a comparison of an RF and AIF that were used as stimuli, as well as a comparison of an AIF generated with the standard image generation vs. an AIF generated with the refined model. Because AI generators are updated quite frequently, it is important to note that all of the refined models were created on October 31st, 2023, and all of the images were generated on November 8th, 2023.

When generating the images, we started with 10 models and eight versions of each possible image were created. To select the final set that were used for the studies outlined below, two independent raters rated them for realism on a 1 to 9 Likert-type scale. Ratings for each image were summed, and the highest rated image from each set was chosen. The original intent was to have 10 images per condition, however for some models, none of the generated images

were deemed realistic (i.e., some images yielded ones from both raters) so these were removed due to their intense lack of realism. Because this did not happen equally across all conditions, in others, several additional low-rated images were removed, resulting in 6 images per condition for the main study.

**Figure 1**

*Real vs. AI-Generated vs. Refined-AI-Generated Images of an Angry, Asian Woman*

*Note.* Images of a real (left panel), AI-generated (middle panel), and refined-AI-generated (right panel) Asian woman with an open-mouthed, angry facial expression. The real image is from the RADIATE database (Conley et al., 2018; Tottenham et al., 2009). The generated images were both created using image-to-image generation on Leonardo.Ai. The refined AI generator was created using the RADIATE images of Asian women with open-mouthed angry expressions. Generation details: Prompt: "an angry asian woman", with a "photography" specification, Guidance Scale = 7, Init Strength = 0.3, eight images were generated and one selected, Prompt Magic off, Alchemy on, Input Dimensions (px) = 768 × 768, using Stable Diffusion v2.1. The base-model AI-generated image was generated on November 28th, 2023; the refined AI-generated image was generated on November 8th, 2023.

### *2.3.2. Participants*

N = 138 undergraduate students ($M_{age}$ = 20.01 years, $SD_{age}$ =2.22 years, $M_{education}$ = 2.36 years, $SD_{education}$ = 1.88 years) from the University of Ottawa's Integrated System for Participation in Research (ISPR) were recruited for this study. Detailed demographic details can be seen in Supplemental Materials Table S1. The total required sample size was estimated using G*Power (Faul et al., 2007) for F-test repeated measures design, within-factors effects, a power of .95 and a small effect size of $f$ = .1 (n = 328) and adjusted using guidelines for repeated measurements (Goulet & Cousineau, 2019). Using the repeated measurements sample size adjustment, the estimated required sample size was N = 108. To do the adjustment, columns of data with no variation among participant responses were removed. Using the remaining data, correlations within each condition were calculated and averaged. Finally, all r-values between conditions were averaged, resulting in one r-value. Using formula 4c in Goulet & Cousineau (2019), where r = .19, m = 6, and $n_m$ = 328, the adjusted sample size was calculated.

Participants were compensated with course credit. This study was approved by the University of Ottawa Social Sciences and Humanities Research Ethics Board.

### *2.3.3. Materials*

As stated above, the stimuli used in the study were images of RFs drawn from the RADIATE Database (Conley et al., 2018; Tottenham et al., 2009) and AIFs created with image-to-image software, respectively. Each image had one of four different facial expressions (angry, fearful, happy, and sad). We had a balanced number of men and women models, and the models were also balanced for race (Asian, Black, and White). All facial expressions were the open-mouthed variants. The total number of stimuli used were 2 (real vs. AI) × 2 (gender) × 3 (race) × 4 (emotion) × 6 (trials/condition) = 288. Randomly distributed within the 288 trials were 4 catch

trials[2]. A demographics questionnaire of our own design and the Prosopagnosia Index (Shah et al., 2015) were also included. As the task involved distinguishing between faces, the Prosopagnosia Index (Shah et al., 2015) helped to identify participants who self-reported having difficulty with facial recognition (n = 3), which aided in our interpretation of results. Scores in the ranges 65-74, 75-84, and 85-100 are broadly indicative of mild, moderate, and severe developmental prosopagnosia, respectively. Because only two participants reached the threshold of mild developmental prosopagnosia and one reached the threshold for moderate prosopagnosia, this scale was not used for analyses or participant exclusion.

#### *2.3.4. Procedure*

The study was conducted online using Qualtrics™. Participants connected online from a location of their own choosing. Participants were first asked to provide informed consent. They then completed the demographics questionnaire, which asked about age, year of study, sex and gender, ethnic background, time spent in Canada, and visual acuity. Next, they completed the Prosopagnosia Index (Shah et al., 2015).

They were then asked to do our main task. This involved viewing 288 images, shown one at a time in random order, and indicating by keypress whether each image was AI-generated or real. Each image was shown on screen until the participant selected a response, followed by a one second inter-stimulus interval. Images were presented at a size of 375 by 375 pixels, which, on the computer in our lab, subtended 10.6 by 10.6 degrees at a viewing distance of 57.3cm.

---

[2] Catch trials consisted of 2 real and 2 AI-generated faces with either “Real” or “AI-generated” typed above them. Prior to the experiment, participants were informed of these catch trials, and were instructed to select the answer that corresponds to the word that appeared above the image when they saw them.

After viewing all the stimuli, participants were asked an open-ended question about what cues they used when determining whether an image was AI-generated or not. Participants were also asked if they were distracted when completing the study, and about their familiarity with AI software. Finally, participants were asked to provide any general comments on the study.

Data has been made available here: https://osf.io/3cd5r/overview?view_only=e07ea639eaa5424aa6ca06556bc41bb7

### *2.3.5. Data Analysis*

For each participant, d' values were calculated for each condition using the following formula, where z is the quantile function assuming a normal standardized distribution, $P_h$ is the proportion of hits, and $P_f$ is the proportion of false alarms.

$$d' = z(P_h) - z(P_f),$$

A response was considered a "hit" if the participant correctly identified an AI-generated images as such. A response was considered a "false alarm" when they incorrectly labelled a real face as AI-generated. Separate d' scores were calculated for each combination of facial expression, gender, and race, for a total of 24 d' scores per participant.

Before further analysis, the data was cleaned using several criteria. This was done in order to maintain good data quality in an online context. Our initial sample size was N = 188. First, participants were removed if they indicated that they needed corrective lenses but were not wearing them (n = 0). Second, participants were removed if they incorrectly answered two or more catch trials (n = 43). Next, outliers were removed based on several criteria. First, statistically outlying d' scores were determined for each condition based on the Van Selst and Jolicoeur's (1994) guidelines. Outliers for completion time were similarly calculated. Participants

were removed if the time it took them to complete the study was an outlier (n = 3). After this, participants with 3 or fewer outlying d' values across the 24 conditions had those values Winsorized, meaning their outlying d’ value was replaced with the highest/lowest non-outlying value in that condition. Participants with four or more outlying d’ values were removed completely (n = 4). This resulted in our final sample size of N = 138.

Following this, a 2 × 3 × 4 repeated measures ANOVA was run with Stimulus Gender (man and woman), Stimulus Race (Asian, Black, and White), and Facial Expression (angry, fearful, happy, and sad) as independent variables. The dependent variable was sensitivity, measured by d’ scores.

In addition to the above, a similar analysis was run examining criterion values per condition. Criterion is a measure from signal detection theory that describes the bias of participants for choosing yes (AI-generated) or no (real). We examined criterion in an exploratory manner, to see if variations in race, gender or expression of face images would have an effect on response biases. Criterion was calculated using the following formula:

$$C = [z(P_h) + z(P_f)]/(-2)$$

A participant is defined as having a lax criterion when they have a tendency to choose "yes" (in this case, "AI-generated"), which is reflected in a negative C-value. A strict criterion is when a participant more often chooses no (in this case "Real"), and their C-value is therefore positive. An unbiased participant will have a criterion at or near zero.

One final exploratory analysis, examining the effects of participant race, was also run. This analysis was carried out in the form of an additional 2 × 2 × 3 × 4 mixed measures ANOVA with Participant Race (BIPOC/Mixed Heritage and White), Stimulus Gender (man and woman),

Stimulus Race (Asian, Black, and White), and Facial Expression (angry, fearful, happy, and sad) as independent variables. The dependent variable was again d'.

## 2.4. Results

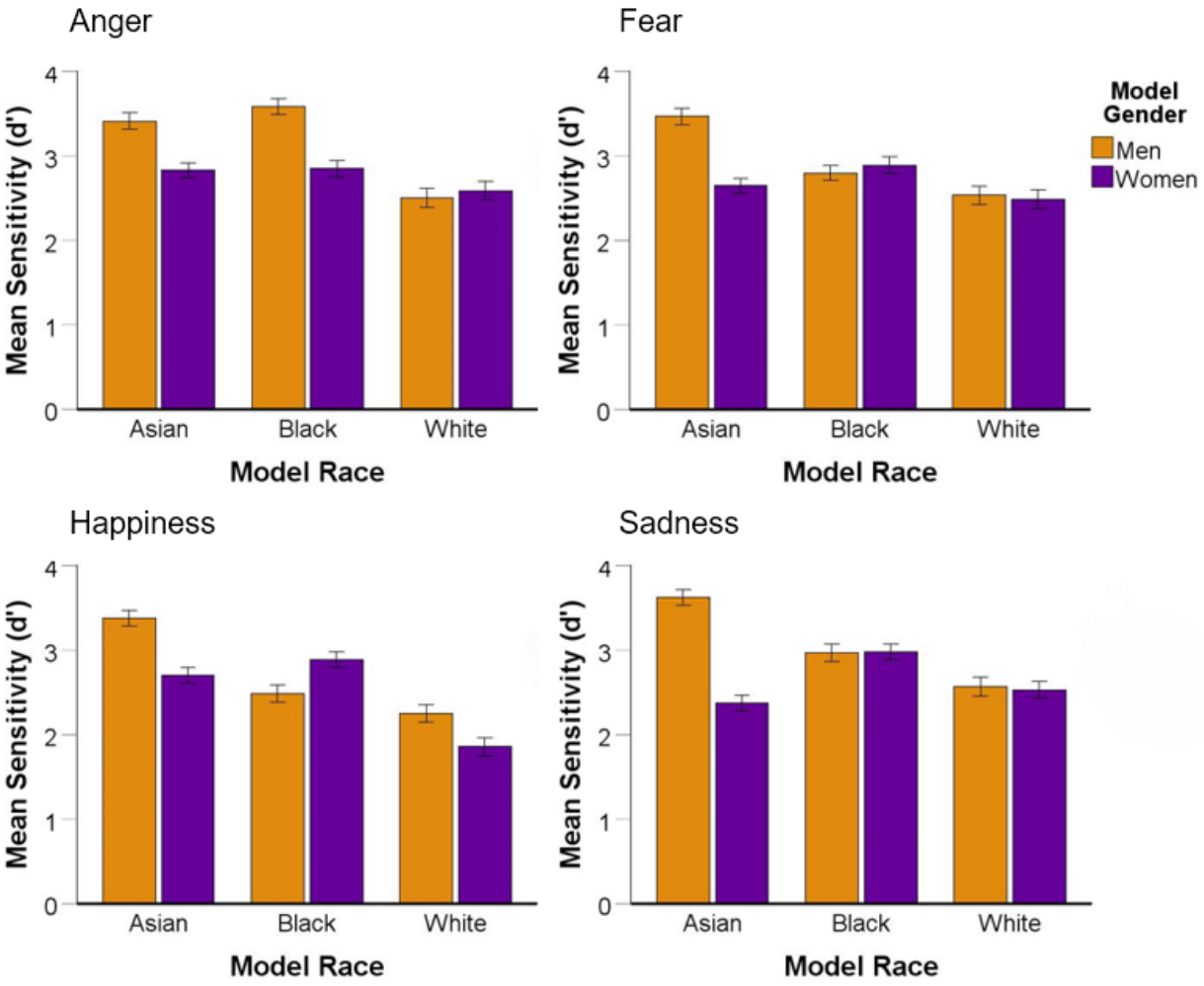
### *2.4.1. Main Analysis: Stimulus Gender × Stimulus Race × Emotional Expression*

Results are illustrated in Figure 2. All pairwise comparisons were Bonferroni adjusted in SPSS. In any case where the assumption of sphericity was violated, Greenhouse-Geisser adjusted values are reported. The main effects of Stimulus Gender [$F(1, 137) = 69.58$, $p < .001$, $\eta_p^2 = .34$], Stimulus Race [$F(1.82, 248.73) = 77.87$, $p < .001$, $\eta_p^2 = .36$], and Emotional Expression [$F(3, 411) = 23.22$, $p < .001$, $\eta_p^2 = .15$] were all statistically significant. The two-way interactions between Stimulus Gender and Stimulus Race [$F(2, 274) = 58.36$, $p < .001$, $\eta_p^2 = .30$], Stimulus Gender and Emotional Expression [$F(3, 411) = 3.00$, $p = .03$, $\eta_p^2 = .02$], and Stimulus Race and Emotional Expression [$F(5.56, 761.54) = 6.91$, $p < .001$, $\eta_p^2 = .05$] were also statistically significant. Finally, the 3-way interaction between Stimulus Gender, Stimulus Race, and Emotional Expression [$F(6, 822) = 14.33$, $p < .001$, $\eta_p^2 = .10$] was statistically significant.

To examine the 3-way interaction, follow-up 2-way repeated measures ANOVAs were run for each of the emotional expressions separately.

**Figure 2**

*Mean Sensitivity (d') to AI-generated Faces*




*Note.* Graphs represent the 2-way interactions between Stimulus Gender and Stimulus Race for each of the following Emotional Expressions: anger (top left), fear (top right), happiness (bottom left), and sadness (bottom right). Error bars represent +/-1 SEM.

#### 2.4.1.1. Anger: Stimulus Gender × Stimulus Race.

The simple main effects of Stimulus Gender [$F(1, 137) = 40.72$, $p < .001$, $\eta_p^2 = .23$] and Stimulus Race [$F(2, 274) = 32.13$, $p < .001$, $\eta_p^2 = .19$] within the Anger level of the emotion

variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2, 274) = 14.89$, $p < .001$, $\eta_p^2 = .10$] was also statistically significant.

Pairwise comparisons revealed that within men's faces, Asian and Black faces had higher d' values than White stimuli [Asian-White: $t(137) = 7.67$, $p < .001$, $d = .65$; Black-White: $t(137) = 8.93$, $p < .001$, $d = .76$]. Moreover, within Asian and Black faces, men's faces had higher d' values compared to women's faces [Asian: $t(137) = 5.05$, $p < .001$, $d = .43$; Black: $t(137) = 6.68$, $p = .001$, $d = .57$].

#### 2.4.1.2. Fear: Stimulus Gender × Stimulus Race.

The simple main effects of Stimulus Gender [$F(1, 137) = 17.54$, $p < .001$, $\eta_p^2 = .13$] and Stimulus Race [$F(1.76, 274.79) = 20.84$, $p < .001$, $\eta_p^2 = .11$] within the Fear level of the emotional expression variable were both statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(1.86, 254.14 = 19.97$, $p < .001$, $\eta_p^2 = .13$] was also statistically significant.

Pairwise comparisons revealed that within men's faces, Asian faces had higher d' values than both Black and White faces [Asian-Black: $t(137) = 6.61$, $p < .001$, $d = .56$; Asian-White: $t(137) = 7.27$, $p < .001$, $d = .61$]. Within women's faces, Black faces had higher d' values than both Asian and White faces [Black-Asian: $t(137) = 2.78$, $p = .02$ , $d = .24$; Black-White: $t(137) = 3.12$, $p = .007$, $d = .27$]. Moreover, only within Asian faces did men's faces have higher d' values than women's [$t(137) = 8.64$, $p < .001$, $d = .74$].

#### 2.4.1.3. Happiness: Stimulus Gender × Simulus Race.

The simple main effects of Stimulus Gender [$F(1, 137) = 9.77$, $p = .002$, $\eta_p^2 = .35$] and Stimulus Race [$F(2,274) = 72.71$, $p < .001$, $\eta_p^2 = .35$] within the Happiness level of the

emotional expression variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2, 274) = 26.29$, $p < .001$, $\eta_p^2 = .16$] was also statistically significant.

Pairwise comparisons revealed that within men's faces, Asian faces had higher d' values than both Black and White faces [Asian-Black: $t(137) = 7.40$, $p < .001$, $d = .63$; Asian-White: $t(137) = 10.88$, $p < .001$, $d = .93$]. Within women's faces, Asian and Black faces had higher d' values than White faces [Asian-White: $t(137) = 7.07$, $p < .001$, $d = .60$; Black-White: $t(137) = 8.31$, $p < .001$, $d = .71$]. When examining the difference between men and women within each race, in both Asian and White faces, men had higher d' values than women [Asian: $t(137) = 6.5$, $p < .001$, $d = .55$; White: $t(137) = 3.36$, $p = .001$, $d = .29$], whereas the opposite was true for Black faces [$t(137) = 3.35$, $p = .001$, $d = .29$].

#### 2.4.1.4. Sadness: Stimulus Gender × Stimulus Race.

The simple main effects of Stimulus Gender [$F(1, 137) = 45.43$, $p < .001$, $\eta_p^2 = .25$] and Stimulus Race [$F(2,274) = 16.84$, $p < .001$, $\eta_p^2 = .11$] within the Sadness level of the emotional expression variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2,274) = 47.64$, $p < .001$, $\eta_p^2 = .26$] was also statistically significant.

Pairwise comparisons revealed that within men's faces, Asian faces had higher d' values than both Black and White stimuli, and Black faces had higher d' values than White faces [Asian-Black: $t(137) = 5.73$, $p < .001$, $d = .49$; Asian-White: $t(137) = 9.12$, $p < .001$, $d = .77$, Black-White: $t(137) = 3.15$, $p = .006$, $d = .27$]. Within women's faces, Black faces had higher d' values than both Asian and White faces [Black-Asian: $t(137) = 5.86$, $p < .001$, $d = .50$; Black-

White: $t(137) = 4.34$, $p < .001$, $d = .37$]. Moreover, only within Asian faces did men's faces have higher d' values than women's faces [$t(137) = 13.00$, $p < .001$, $d = 1.11$].

### *2.4.2. Exploratory Analyses*

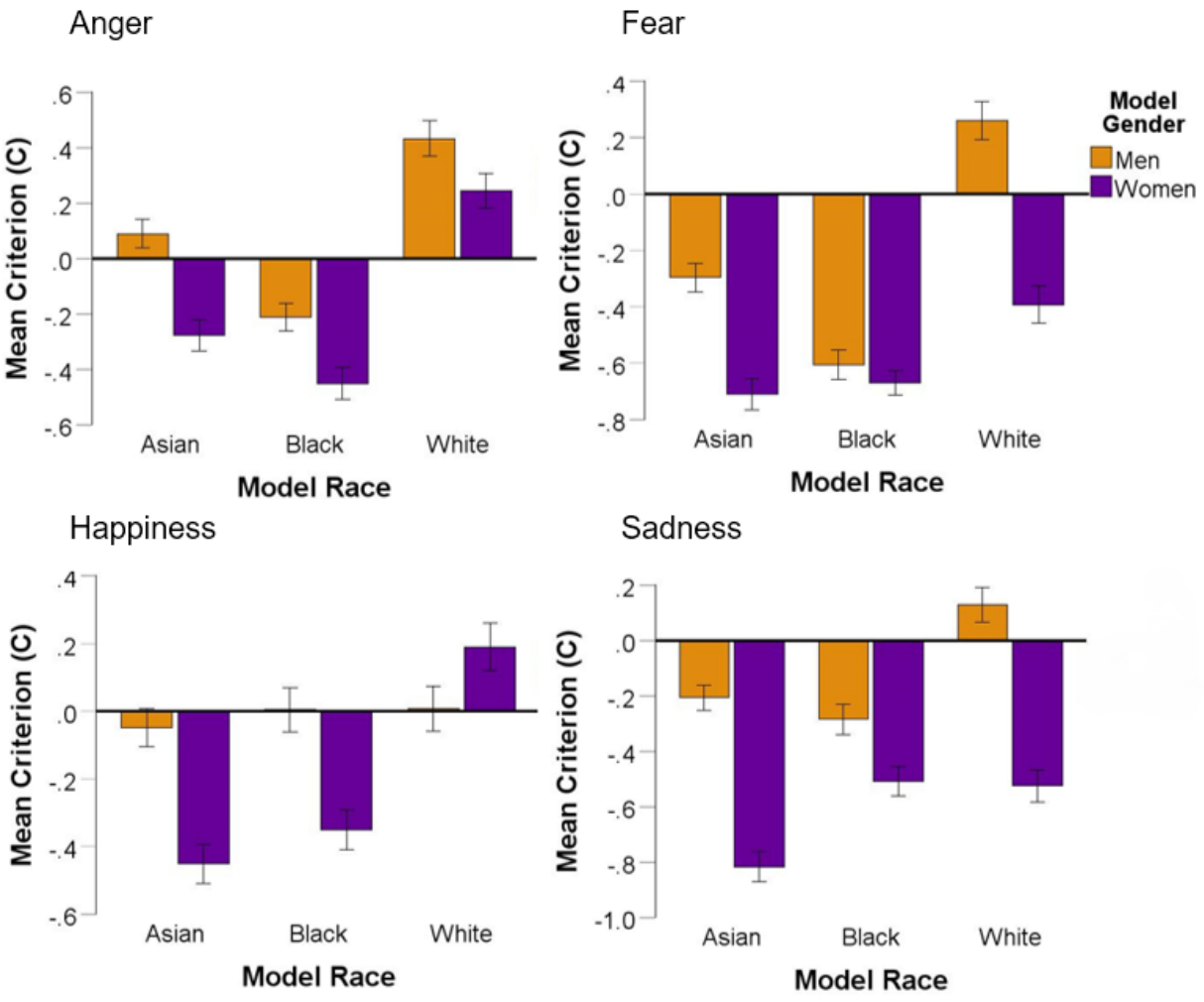
#### 2.4.2.1. Criterion.

Results are illustrated in Figure 3. As an exploratory analysis, Criterion values were examined in the same way as sensitivity. That is, a 2 × 3 × 4 repeated measures ANOVA with Stimulus Gender, Stimulus Race, and Emotional Expression as independent variables was run. Where the assumption of sphericity was violated, Greenhouse-Geisser adjusted values are reported. The main effects of Stimulus Gender [$F(1, 137) = 234.06$, $p < .001$, $\eta_p^2 = .63$], Stimulus Race [$F(1.80, 246.15) = 101.34$, $p < .001$, $\eta_p^2 = .43$], and Emotional Expression [$F(2.70, 370.23) = 88.69$, $p < .001$, $\eta_p^2 = .39$] were all statistically significant. The two-way interactions between Stimulus Gender and Stimulus Race [$F(2, 274) = 13.20$, $p < .001$, $\eta_p^2 = .09$], Stimulus Gender and Emotional Expression [$F(3, 411) = 17.18$, $p = .03$, $\eta_p^2 = .11$], and Stimulus Race and Emotional Expression [$F(5.38, 737.09) = 19.54$, $p < .001$, $\eta_p^2 = .13$] were likewise all statistically significant. Finally, the 3-way interaction between Stimulus Gender, Stimulus Race, and Emotional Expression [$F(6, 822) = 23.43$, $p < .001$, $\eta_p^2 = .15$] was statistically significant.

To further examine the significant 3-way interaction, separate 2-way ANOVAs were run within each emotional expression condition. Overall, participants more often classified images of people of colour and women as AI-generated, compared to White men being classified more often as real.

**Figure 3**

*Mean Criterion (C) when Choosing Between AI-Generated and Real*




*Note.* Graphs represent the 2-way interactions between Stimulus Gender and Stimulus Race for each of the following Emotional Expressions: anger (top left), fear (top right), happiness (bottom left), and sadness (bottom right). Error bars represent +/-1SEM. In this case, a negative Criterion indicates a bias towards choosing AI-generated, while a positive Criterion indicates a bias towards choose Real.

#### *2.4.2.1.1. Anger: Stimulus Gender × Stimulus Race.*

The simple main effects of Stimulus Gender [$F(1, 137) = 49.83$, $p < .001$, $\eta_p^2 = .27$] and Stimulus Race [$F(1.87, 255.94) = 117.25$, $p < .001$, $\eta_p^2 = .46$] within the Anger level of the emotional expression variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2, 274) = 3.06$, $p = .05$, $\eta_p^2 = .02$] was also statistically significant.

Comparing within Stimulus Gender, all pairwise comparisons were significant at $p < .001$ with *t*-values ranging from 5.55 to 9.92, and Cohen's *d* ranging from .47 to .84, aside from the difference between Asian women and Black women [$t(137) = 3$, $p = .011$, $d = .26$]. Comparing within Stimulus Race, all pairwise comparisons were statistically significant at $p < .001$ (*t*-values ranging from 4.80 to 6.46 and Cohen's *d* ranging from .41 to .55) aside from the comparison of White men and White women [$t(137) = 2.93$, $p = .004$, $d = .25$].

#### *2.4.2.1.2. Fear: Stimulus Gender × Stimulus Race.*

The simple main effects of Stimulus Gender [$F(1, 137) = 121.78$, $p < .001$, $\eta_p^2 = .47$] and Stimulus Race [$F(1.84, 252.35) = 78.07$, $p < .001$, $\eta_p^2 = .36$] within the Fear level of the emotional expression variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2, 274) = 26.82$, $p < .001$, $\eta_p^2 = .16$] was also statistically significant.

Comparing within Stimulus Gender, all pairwise comparisons were significant at $p < .001$ with *t*-values ranging from 4.40 to 12.20, and Cohen's *d* ranging from .37 to 1.04, aside from the difference between Asian women and Black women ($p = 1.00$). Comparing within Stimulus Race, Asian men had higher Criterion values compared to Asian women [$t(137) = 8.43$, $p < .001$, $d = .72$], and White men had higher Criterion values compare to White women [$t(137) = 10.20$, $p < .001$, $d = .87$].

#### *2.4.2.1.3. Happiness: Stimulus Gender × Stimulus Race.*

The simple main effects of Stimulus Gender [$F(1, 137) = 30.86$, $p < .001$, $\eta_p^2 = .18$] and Stimulus Race [$F(1.89, 259.37) = 28.45$, $p < .001$, $\eta_p^2 = .17$] within the Happiness level of the emotional expression variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2, 274) = 28.91$, $p < .001$, $\eta_p^2 = .17$] was also statistically significant.

Comparing within Stimulus Gender, there were no differences between Asian, Black, and White men ($p = 1.00$). However, White women had higher Criterion values than both Asian and Black women [White-Asian: $t(137) = 10.34$, $p < .001$, $d = .88$; White-Black: $t(137) = 8.07$, $p < .001$, $d = .69$]. Comparing within Stimulus Race, Asian and Black men had higher Criterion values compared to Asian and Black women, respectively [Asian: $t(137) = 6.71$, $p < .001$, $d = .57$; Black: $t(137) = 5.84$, $p < .001$, $d = .50$]. Whereas White women had higher Criterion values than White men [$t(137) = 3.05$, $p = .003$, $d = .26$].

#### *2.4.2.1.4. Sadness: Stimulus Gender × Stimulus Race.*

The simple main effects of Stimulus Gender [$F(1, 137) = 203.61$, $p < .001$, $\eta_p^2 = .60$] and Stimulus Race [$F(2,274) = 24.15$, $p < .001$, $\eta_p^2 = .15$] within the Sadness level of the emotional expression variable were statistically significant. The interaction between Stimulus Gender and Stimulus Race [$F(2, 274) = 18.63$, $p < .001$, $\eta_p^2 = .12$] was also statistically significant.

Comparing within Stimulus Gender, Asian and Black men had lower Criterion values than White men [Asian-White: $t(137) = 5.33$, $p < .001$, $d = .45$; Black-White: $t(137) = 6.45$, $p < .001$, $d = .55$]. However, Asian women had lower Criterion values than both Black and White women [Asian-Black: $t(137) = 5.62$, $p < .001$, $d = .48$; Asian-White: $t(137) = 4.87$, $p < .001$, $d$

= .41]. Comparing within Stimulus Race, all pairwise comparisons were statistically significant at $p < .001$ ($t$-value ranging from 4.02 to 11.31, and Cohen's $d$ ranging from .34 to .96).

#### 2.4.2.2. Participant Race.

To examine the effects of participant race, a 2 × 3 × 2 × 4 mixed measures ANOVA with Stimulus Gender (man and woman), Stimulus Race (Asian, Black, and White), Participant Race (BIPOC/Mixed Heritage and White), and Facial Expression (angry, fearful, happy, and sad) as independent variables, and d' as the dependent variable. This analysis revealed no statistically significant effects relating to participant race (all $p > .05$).

### 2.5. Discussion

The goal of the present study was to examine biases in the outputs of AI image generators as they created faces with different races, genders, and facial expressions. Participants were shown faces under 24 conditions and asked to classify them as either real or AI-generated. The conditions were as follows: stimulus gender (man and woman), stimulus race (Asian, Black, and White), and emotional expression (anger, fear, happiness, and sadness). Half of the images in each condition were photographs and half were AI-generated images. Several methodological elements were implemented to improve the validity of the study. First, for the real faces, we used a validated set of faces (RADIATE; Conley et al., 2019, Tottenham et al., 2009). Second, using this database, we refined generative models to weight them towards our different conditions, which should have lessened the bias in the initial training set of the generator; In this way, we gave the AI generators the best chance at producing equally convincing faces across the different conditions. Finally, we used image-to-image generation, using the images from the RADIATE database, which should have further helped to reduce bias from the initial training set images.

Two predictions and two research questions were made based on previous work in this area. We discuss below the results and implications for each hypothesis and research question.

***H1:*** *Because AI generators are trained disproportionately on White face stimuli (Kärkkäinen & Joo, 2021), they will be biased towards generating more realistic White faces compared to faces of people of colour. We therefore predict that AI Faces (AIFs) of people of colour will be more accurately distinguished from real faces than will AIFs of White people.*

Results support hypothesis 1, AIFs of people of colour were more detected, compared to AIFs of White faces. This finding supports the idea that AI generators are biased towards generating more realistic White faces, likely due to bias in the training images used (Kärkkäinen & Joo, 2021). Overall, faces of people of colour were more accurately classified as either real or AI-generated, but there were some variations. Broadly speaking, AIFs of Asian men were more accurately detected across emotions. However, AIFs of Black men were only more accurately identified in the Anger condition, while those of Black women were more accurately identified in the Happy and Sad conditions.

These findings are in line with previous research, which found an overall advantage at detecting AIFs of people of colour, specifically of Asian individuals (Nightingale & Farid, 2022). Moreover, this is in line with recent research which finds that AI-generators create more stereotyped images of people of colour and women, compared to White men (AlDahoul et al., 2025; Assis & Moura, 2025; Bianchi et al., 2023 Chauhan et al., 2024; Gawali et al., 2024; Ghosh & Caliskan, 2023; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Mubashir, 2024; Park, 2024). This indicates that not only are White faces disproportionately represented in AI training sets, but that representation of different non-white races is not equal. This implies that improving the ability of AI generators to create realistic faces across the range of human

physiognomies will require training sets to include images from a broad and diverse range of faces.

***H2:*** *Because AI generators are trained disproportionately on images of men's faces (Kärkkäinen & Joo, 2021), they will be biased towards generating more realistic faces of men compared to faces of women. We therefore predict that AIFs of women will be more accurately distinguished from real faces than will AIFs of men.*

Results do not support hypothesis 2. Indeed, contrary to the hypothesis, and previous research (Assis & Moura, 2025; Gorska & Jemielniak, 2023; Locke & Hogdon, 2024; Mubashir, 2024, Wu et al., 2025), men's faces were either *more* accurately categorized as AI-generated or Real than women's faces or were identified at an equal rate to women's faces. This was the case in all but one condition. Faces of happy Black women were more detected than faces of happy Black men. Collapsed across all other factors, men's faces were more accurately categorized than those of women. This, however, must be qualified by the significant 3-way interaction. In all emotional expression conditions, AIFs of Asian men were more accurately detected than those of Asian women. However, for Black faces, this difference is only seen in the angry faces, and the opposite is true for happy faces, where AIFs of Black women were more accurately identified as compared to those of Black men. Finally, for White faces, the only significant difference is seen with happy faces, where AIFs of men were more accurately detected than those of women.

The current study's results clearly do not support the hypothesis that male faces are over-represented in the training set for the AI generator. Rather, the results seem to indicate an over-representation of women's faces. However, based on the criterion analysis, these results should be interpreted with caution. When looking at criterion values, women's faces, more specifically

women of colour, were more often labelled as AI-generated. This was true across all emotional expressions. This may indicate that participants were getting more false alarms for the women's faces. Therefore, it may not be that women's faces are over-represented in AI generator training sets, leading to more realistic faces, but rather, the real faces we used, for reasons unknown to us, were perceived as AI-generated more often than the real faces of men.

***RQ1:*** *Will AIFs with certain facial expressions be less accurately distinguished than AIFs with other facial expressions? It is possible that the stimulus training sets of AI generators contain disproportionately more of certain expressions than others (e.g., more smiling faces as compared to angry, fearful, and sad ones). This would potentially cause them to generate more realistic images of certain expressions compared to others.*

Results indicate that emotional expression does impact participants' ability to distinguish the AI-generated faces from real faces. This is supported by the main effect of emotional expression, where angry faces were more accurately identified compared to fearful and happy faces; and fearful and sad faces were more accurately identified compared to happy faces. This is also supported by the 3-way interaction, which shows a different pattern of identification across races and genders for each of the emotional expressions. However, in the current sample, there is no clear pattern as to how emotional expression impacts participants' ability to distinguish between real and AI-generated faces. More research is therefore needed into the impacts of emotional expression.

***RQ2:*** *Will the factors of race, gender, and expression interact in determining the verisimilitude of AIFs? The lack of research in this area makes it difficult to form firm hypotheses about interactions of the above effects. However, we expect that they may interact with each other such that, for example, an AIF of a sad woman of colour will be more accurately*

*distinguished than an AIF of a happy White man. However, it is not clear to what degree the different factors (race, gender, expression) would interact rather than producing additive effects.*

As previously discussed, there was a significant 3-way interaction between stimulus gender, stimulus race, and emotional expression. To reiterate, this interaction is likely due to the different patterns of sensitivity across the emotional expressions. This seems to indicate that the factors interact to determine the verisimilitude of AIFs. Specifically, even though men's faces, overall, were more accurately identified as AI generated vs. real, AIFs of Asian men were even more accurately identified than those of White or Black men's faces, and this was the case across all emotional expressions. In contrast, identification rates varied across emotional expressions for the other gender and race conditions. This supports the idea that AI generators are biased in terms of race and gender when creating realistic images of faces. Specifically, they may be biased against creating realistic faces of Asian men, angry Black men, happy Black women, and sad Black men and women.

**2.6. Limitations**

As previously stated, this area of research is quite new and is evolving rapidly. AI generators are advancing, and new ones are being released frequently, and so the state of AI-generated images is constantly evolving. As this is the case, a limitation of most research in this area is that the faces being used may become quickly outdated, and the research will need to be replicated with newer versions of the software.

Moreover, not all generators will give the same results. For example, one study (Fraser et al., 2023) examined differences in gender and race biases for three different AI generators. Because we used Leonardo.ai, and specifically our own refined models, the results may not be as generalizable to other generators. However, this limitation was weighed among pros and cons of

generating the images in this way, and we believe that at the time of stimulus creation, using the refined models tool on Leonardo.ai was one way to get the most realistic possible faces across our many conditions.

The large proportion of women to men meant we couldn't examine the impacts of participant gender. This is often an issue when using a convenience sample of undergraduate students. Future research should replicate the study using a community sample with a more representative proportion of men and women.

Finally, our sample size, while large enough to detect significant differences in our main analysis, may not have been sufficient to detect differences in the exploratory analysis of participant race. Moreover, to ensure equal groups for this analysis, participant race had to be collapsed into only two groups (White and BIPOC/Mixed Heritage), rather than grouping participants by more refined groups. Future research should examine the differences between different racial groups when examining stimuli of difference races.

### 2.7. Conclusion

Our results suggest that there is a demographic bias in AI generators. Although AI-generated faces overall were accurately identified, there was a clear difference in the accuracy for White faces compared to Asian and Black faces. The bias in terms of gender is less clear, as the faces of men were more accurately distinguished than faces of women, specifically for Asian individuals. However, when taken with the criterion analysis, showing that women overall were classified as AI-generated, indicating a higher rate of false alarms, this result does not point as clearly to a bias in either direction. More research is needed to confirm how stimulus gender impacts accuracy in distinguishing AI-generated faces from real faces. With regards to the exploratory results, emotional expression does seem to impact identification of AIFs vs. RFs,

however there was no clear pattern of results. More research is therefore needed to examine the effect of emotional expression. Stimulus gender, stimulus race, and emotional expression do seem interact, especially stimulus gender and stimulus race. More research is needed to confirm the findings; however, our results show preliminary support for the idea that demographic characteristics interact when examining bias in AI generators. Finally, with regards to the exploratory results of participant race, no significant effects were found, indicating that participant race does not impact identification of AIFs vs. RFs, however the usual caveats when interpreting null results apply. In conclusion, in the training and refinement of AI generators, developers should be mindful of the fact that certain groups may be over-represented in the existing training sets, and those under-represented may not be under-represented to the same extent. With the increasing prevalence of AI-generated images in our daily lives, how different genders and races are represented may have important implications for how we view individuals belonging to these groups.

**2.8. References**

AlDahoul, N., Rahwan, T., & Zaki, Y. (2025). *AI-generated faces influence gender stereotypes and racial homogenization.* Sci Rep 15(1):14449. https://doi.org/10.1038/s41598-025-99623-3

Assis, J. D., & Moura, M. A. (2025). Algorithmic Semiosis and Racial Bias: A Study of Images Created by Generative AI. *Encontros Bibli*, *30*, e103495. DOI: 10.5007/1518-2924.2025.e103495

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. https://doi.org/10.1145/3593013.3594095

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake'images of human faces. *Journal of Cybersecurity*, *9*(1), tyad011.

Chandaliya, P. K., & Nain, N. (2022). ChildGAN: Face aging and rejuvenation to find missing children. *Pattern Recognition*, *129*, 108761. https://doi.org/10.1016/j.patcog.2022.108761

Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., & Vanga, R. (2024, February). Identifying race and gender bias in diffusion ai image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)* (pp. 1-6). IEEE. DOI: 10.1109/ICAIC60265.2024.10433840

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, *270*, 1059–1067. https://doi.org/10.1016/j.psychres.2018.04.066

Dhariwal P, Nichol A (2021) Diffusion models beat GANs on image synthesis. Adv Neural Inf Process Syst 34:8780–8794

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

FBI (Federal Bureau of Investigation). (2023, June 5). *Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. https://www.ic3.gov/Media/Y2023/PSA230605

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2023). *A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes when the Input is Under-Specified?* (arXiv:2302.07159). arXiv. http://arxiv.org/abs/2302.07159

Gawali, A., Tarle, A., Can Inac, C., & Kulkarni, S. S. (2024). A Framework for Detecting Stereotypes, Prejudices and Discrimination in AI-Generated Imagery. ResearchGate. https://doi.org/10.13140/RG.2.2.17901.29926

Ghosh, S., & Caliskan, A. (2023). 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6971–6985. https://doi.org/10.18653/v1/2023.findings-emnlp.465

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144. https://doi.org/10.1145/3422622

Gorska, A. M., & Jemielniak, D. (2023). The invisible women: uncovering gender bias in AI-generated images of professionals. *Feminist Media Studies*, *23*(8), 4370-4375. https://doi.org/10.1080/14680777.2023.2263659

Goulet, M.-A., & Cousineau, D. (2019). The Power of Replicated Measures to Increase Statistical Power. *Advances in Methods and Practices in Psychological Science*, *2*(3), 199–213. https://doi.org/10.1177/2515245919849434

*Home | Leonardo.Ai*. (n.d.). Leonardo.Ai. https://app.leonardo.ai/

Jääskeläinen, P., Sharma, N. K., Pallett, H., & Åsberg, C. (2025). Intersectional analysis of visual generative AI: the case of stable diffusion. *AI & SOCIETY*, 1-22. https://doi.org/10.1007/s00146-025-02207-y

Josephs, E., Fosco, C., & Oliva, A. (2023). Artifact magnification on deepfake videos increases human detection and subjective confidence. J vis (Charlottesville, Va) *23*(9):5327. https://doi.org/10.1167/jov.23.9.5327

Kärkkäinen, K., & Joo, J. (2021). *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age*. Proceedings/IEEE Workshop on Applications of Computer Vision, 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159

Locke, L. G., & Hodgdon, G. (2024). Gender bias in visual generative artificial intelligence systems and the socialization of AI. *AI & SOCIETY*, 1-8. https://doi.org/10.1007/s00146-024-02129-1

Moshel, M. L., Robinson, A. K., Carlson, T. A., & Grootswagers, T. (2022). Are you for real? Decoding realistic AI-generated faces from neural activity. *Vision Research*. https://doi.org/10.1016/j.visres.2022.108079

Mubashir, M. (2024). The gendered dress of DALL-E 2: Exploring profession-based images in the Indian context. *MedieKultur: Journal of media and communication research*, *40*(76), 100-119. https://doi.org/10.7146/mk.v40i76.143565

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.

Park, Y. S. (2024). White default: Examining racialized biases behind AI-generated images. *Art Education*, *77*(4), 36-45. https://doi.org/10.1080/00043125.2024.2330340

Rossi, S., Kwon, Y., Auglend, O. H., Mukkamala, R. R., Rossi, M., & Thatcher, J. (2023). Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles? *Proceedings of the 56th Hawaii International Conference on System Sciences,* 134–143. https://hdl.handle.net/10125/102645. Accessed 16 Sept 2025.

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343.

American Psychological Association, APA Task Force on Race and Ethnicity Guidelines in Psychology. (2019). Race and Ethnicity Guidelines in Psychology: Promoting Responsiveness and Equity. Retrieved from http://www.apa.org/about/policy/race-and-ethnicity-in-psychology.pdf

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

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

Van Selst, M., & Jolicoeur, P. (1994). A Solution to the Effect of Sample Size on Outlier Elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. https://doi.org/10.1080/14640749408401131

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F.-Y. (2017). Generative adversarial networks: Introduction and outlook. IEEE/CAA Journal of Automatica Sinica, 4(4), 588–598. IEEE/CAA Journal of Automatica Sinica. https://doi.org/10.1109/JAS.2017.7510583

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": The mad world of AI-generated media. *IT Professional*, *22*(5), 90–99.

Wu, Y., Nakashima, Y., & Garcia, N. (2025). Revealing gender bias from prompt to image in stable diffusion. *Journal of Imaging*, *11*(2), 35. doi: 10.3390/jimaging11020035

## 2.9. Supplemental Materials

**Table S1**

***Demographic Details of Participants***

| | Mean | SD |
|---|---|---|
| Age (Years) | 20.01 | 2.22 |
| Education (Years) | 2.36 | 1.88 |
| Time spent in Canada (Years) | 10.38 | 5.27 |
| | N | Percentage (%) |
| Race | | |
| Arab | 17 | 12.3 |
| Arab/African | 10 | 7.2 |
| Asian-East | 12 | 8.7 |
| Asian-South | 10 | 7.2 |
| Asian-Southeast | 4 | 2.9 |
| Black-African | 15 | 10.9 |
| Black-Caribbean | 2 | 1.4 |
| Black-North American | 1 | .7 |
| First Nation | 1 | .7 |
| Indian-Caribbean | 2 | 1.4 |
| Latin American | 3 | 2.2 |
| Metis | 2 | 1.4 |
| Middle Eastern | 6 | 4.3 |
| Mixed Heritage | 6 | 4.3 |

| | | |
|---|---|---|
| White-North American | 50 | 36.2 |
| White-European | 22 | 15.9 |
| You don't have an option that applies to me | 1 | .7 |
| Choose not to respond | 2 | 1.4 |
| Gender | | |
| Agender | 2 | 1.4 |
| Genderqueer | 1 | .7 |
| Man | 22 | 15.9 |
| Non-binary | 2 | 1.4 |
| Woman | 112 | 81.2 |
| Questioning | 1 | .7 |
| Choose not to respond | 1 | .7 |

*Note.* Mean and Standard Deviation (SD) for age (n = 137), years of education (n = 136), and time spent in Canada for participants. For time spent in Canada, only those who reported living somewhere else were included (n = 37) Number of participants (N) and percentage of race and gender of participants (N = 138). The total percentage for race and gender is higher than 100% as participants were instructed to select all that apply, so some participants selected more than one option. Although Mixed Heritage was an option, not all participants who selected multiple options selected this, and not all participants who selected Mixed Heritage specified further.

Study 2 has been submitted for publication.

## 3. STUDY 2: Recognition of Emotional Expressions in Real vs. AI-generated Face Images: Effects of Race, Gender, and Expression

Megan Lawrence* & Charles A. Collin

### 3.1. Abstract

Emotion recognition is vital to everyday interactions. Artificial intelligence (AI) generated images of faces are rapidly becoming ubiquitous. Previous work shows that our ability to detect these images is declining, leading to a range of possible negative outcomes. We therefore need a better understanding of how and when we can detect AI-generated faces. In this study we examined whether emotions expressed in AI-generated faces are as intense, salient, and accurately identified as those of real faces. N = 450 undergraduate students rated a mix of real and AI-generated faces on seven emotions. Faces varied by gender, race, and expression. Emotional expressions of AI-generated faces were generally rated as less intense and salient, and were less often accurately identified, compared to those of real faces. Interaction effects with model race and gender showed that AI image generators reflect various societal biases and stereotypes when asked to create faces of different genders and races.

### 3.2. Introduction

Emotion recognition guides our behaviours towards others. We are constantly decoding the emotions of other people and making inferences about them based on their emotional state. In turn, this impacts how we behave towards others (Van Kleef, 2009). Therefore, the decoding of emotions is vital in everyday life. Whether the same is true when viewing AI-generated images is unclear.

In real faces, emotional expression recognition has been studied extensively. Of particular relevance for the present work are previous studies on sociodemographic factors, such as the race[3] and gender of the participants and the stimulus faces. For example, several studies have found an advantage for female participants in emotion recognition (Kret & De Gelder, 2012; McClure, 2000; Montagne et al., 2005). Other studies have examined the effects of observer and model race on emotion recognition. Although there does seem to be an own-race advantage in some situations, there is not an overall advantage for own-race faces (Beaupré & Hess, 2005; Campbell et al., 2010; Jiang et al., 2023; Young et al., 2011). For example, Jiang et al. (2023) found that European Caucasian participants had lower overall accuracy for emotion recognition of other-race faces, while East Asian participants did not. More specifically, both groups showed lower accuracy for angry other-race faces, and both groups showed lower accuracy for sad East Asian faces, though the effect was weaker for East Asian participants. Of note in this study was that they recruited only female participants. To the extent that the effects of perceiver race might interact with those of perceiver gender, this makes their results less

[3] The term race is used throughout based on APA guidelines defining race as "social construction and categorization of people based on perceived shared physical traits that result in the maintenance of a sociopolitical hierarchy" (American Psychological Association, 2019, p. 47).

generalizable. More generally, some authors have found that certain emotions are easily confused. For example, anger and disgust, and fear and surprise are often confused for one another (Wang et al., 2019), but the interaction of this effect with socio-demographic factors has not been extensively studied.

Research with AI-generated images is less extensive. One recent study (Eiserbeck et al., 2023) used EEG measurements and subjective ratings of real faces under *presumed real* and *presumed fake* conditions; This meant that, although all the faces were of real individuals, in the presumed fake condition, participants were told that the faces were AI-generated. They found that presumed fake positive faces had reduced intensity, slower evaluations, and more effortful evaluations. Relatedly, some recent studies have examined certain social perceptions of AI-generated face images (Liefooghe et al., 2023; Nightingale & Farid, 2022; Tucciarelli et al., 2022). For example, Nightingale and Farid (2022) found that trustworthiness ratings of AI-generated faces were *higher* than those of real faces. Moreover, they found small effects of model race on trustworthiness ratings, whereby Black faces were rated as slightly more trustworthy than South Asian faces. Moreover, female faces were rated as more trustworthy overall compared to male ones (Nightingale & Farid, 2022). Similarly, Liefooghe et al. (2023) found that when participants rated faces labelled as real, they found them more trustworthy than those labelled as AI-generated, even though all the faces were real. In contrast, Tucciarelli et al. (2022) found higher social conformity to AI-generated faces as compared to real ones. Moreover, participants conformed more to faces they had previously rated as real, regardless of whether they actually were real.

Research on perceptions of AI-generated faces has primarily focused on social perceptions, rather than emotions. Our judgements of emotions of others influence the inferences

we make about them and guide our social behaviours. AI-generated faces are already being used in a wide variety of contexts, such as in movies and restorations of low-quality data (Eqbal, 2023; Whittaker et al., 2020), and there is potential to use them in research. Knowing whether we can accurately identify emotions in AI-generated faces will aid in understanding how effective they are in these areas, especially as they become more commonplace. The current study aimed to do this by examining the differences in accuracy of emotion detection, the saliency of emotions, and the intensity of emotions between real and AI-generated faces across different races and genders.

One concern here is the comparability of the AI-generated and real faces. It could be that AI-generated faces are overall more or less intense in their emotional expressions than real ones. To overcome this, one could engage in extensive prompt engineering and/or generate a large set of faces and select the ones most comparable to real faces. But our interest here is in examining the interaction between the source of the faces (i.e., photographs vs. AI-generated) and demographic factors (race and gender). To that end, we used the same simple prompt for all faces images, only changing the requested race, gender, and expression of the face. This enabled us to answer the question: If we apply the same conditions to generating faces of different races, genders, and emotional expressions, will there be differences between them, and will there be differences in how they compare to their real face counterparts?

Previous work has shown that AI-generators are biased towards generating more realistic faces of White men compared to other races and women. Previous work (Lawrence, Cimermanis, & Collin, 2025) showed compatible differences in ability to discriminate AI-generated images from real ones based on model gender, race, and emotional expression. Furthermore, previous research has shown that AI-generators may have gender and racial biases that emerge when one

attempts to generate inclusive, realistic images (AlDahoul et al., 2025; Bianchi et al., 2023; Ghosh & Caliskan, 2023; Kärkkäinen & Joo, 2021). Since the time of data collection, recent work has also demonstrated a bias when generating images of women and people of colour, and a bias against them when using generic prompts (AlDahoul et al., 2025; Assis & Moura, 2025; Bhardwaj et al., 2025; Gawali et al., 2024; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Locke & Hogdon, 2024; Mubashir, 2024; Park, 2024; Wu et al., 2025). Therefore, it is plausible that biases may also exist when trying to generate an inclusive set of faces with varying emotional expressions. Moreover, it is possible that the degree of bias will vary with the emotional expression, perhaps reflecting societal perceptions of different groups. Some authors have attempted to create datasets that are more inclusive, for the purpose of training AI generators (Kärkkäinen & Joo, 2021).

Based on the above research with real faces, and the few studies on AIFs, we pre-registered our hypotheses, which can be found here: https://osf.io/dqvm4/overview?view_only=439175e4c34642f1a718054b12213c57. The hypotheses are reiterated hereunder:

**H1:** Emotion recognition will be poorer with AI-generated faces than real faces, particularly for the AI-generated faces of people of colour compared to the White AIFs.

**H2:** Emotion intensity ratings will be lower for AI-generated faces than real faces, particularly for the AI-generated faces of people of colour compared to the White AIFs.

**H3:** Saliency measures will be lower for AI-generated faces than real faces, particularly for the AI-generated faces of people of colour compared to the White AIFs.

**H4:** Happy faces will yield the best results compared to the other emotional expressions, such that they will yield better emotion recognition, higher intensity, and higher saliency.

As the research in this area is quite sparse, it was difficult to form firm hypotheses regarding the effects of stimulus characteristics. Moreover, it is difficult to predict how emotional expression may interact with stimulus characteristics. Therefore, the following question was used to guide exploration of how these factors might interact in determining the levels of our measures of facial expression perception: Will the outcome measures differ based on stimulus gender?

We expected that the above effects might interact with each other in complex ways. For example, an AI-generated image of a sad, Black woman might be less accurately identified as sad, less emotionally salient, and less intense than an AI-generated image of a happy, White man. However, it was not clear *a priori* whether the different factors (race, gender, expression) would interact rather than producing additive effects.

## 3.3. Methods

### *3.3.1. Participants*

N = 450 undergraduate students from the University of Ottawa's Integrated System for Participation in Research (ISPR), an undergraduate subject pool, were recruited for this study (See Supplemental Materials Table S1 for demographic breakdown). The total required sample size was estimated using GPower (Faul et al., 2007) for F-test repeated measures design, within factors effect, a power of .95 and a small effect size of $f = 0.1$ (n = 328). Using Goulet and Cousineau's (2019) repeated measurements sample size adjustment, the estimated sample size was N = 302. We used equation 4c in their paper, with r = .88, $n_1 = 328$, and m = 3 (Goulet & Cousineau, 2019). Our sample size of N = 450 was sufficient to detect a small effect size of *f*

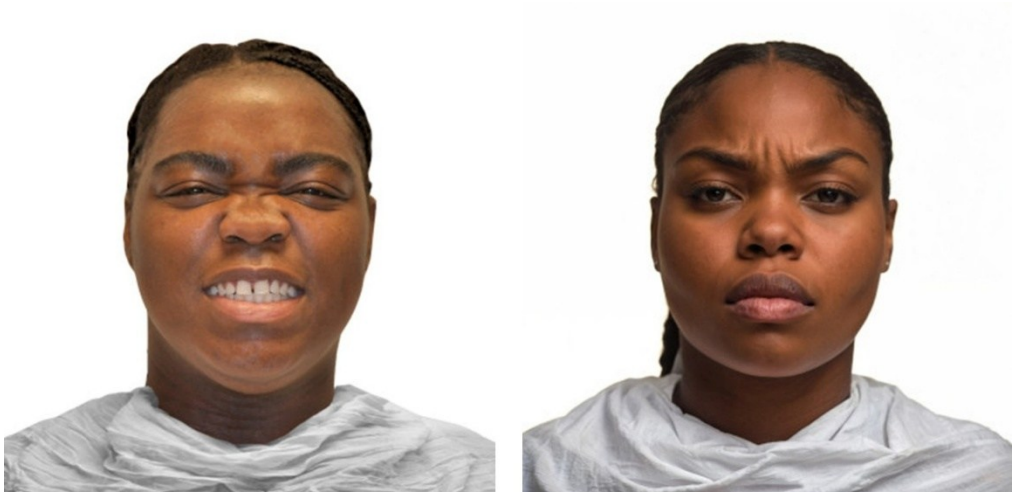
= .09 with 95% probability. Participants were compensated with course credit. This study was approved by the University of Ottawa Social Sciences and Humanities Research Ethics Board.

### *3.3.2. Materials*

At the time of stimulus creation (February 24, 2024), Leonardo.ai had released a photorealistic model for generating images. We decided on the following prompt: "a 20 to 30 year old angry/scared/happy/sad asian/black/white woman/man with a white sheet covering her/his shoulders" and used image-to-image software to generate AI-generated faces from real faces from the RADIATE database (Conley et al., 2018; Tottenham et al., 2009; See Figure 1 for an example) using the photorealistic model.

**Figure 1**

*Example of an AI-generated Face Using the Photorealistic Model Compared to a Real Face*



*Note.* Images of a Black woman with an angry facial expression. The real image (left) was drawn from the RADIATE database (Conley et al., 2018; Tottenham et al., 2009), and the AI-generated image (right) was created using image-to-image generation on Leonardo.Ai with images from the RADIATE database. Generation details: Model: Photorealistic, Leonardo Diffusion XL, HDR; Prompt: "a 20 to 30 year old angry black woman with a white sheet covering her shoulders"; Guidance Scale = 7; Strength = 0.45; eight images

were generated and one selected, Alchemy on, Input Dimensions (px) = 768 × 768. The AI-generated image was generated on February 24th, 2024.

After stimulus creation, the images were validated in several steps. Detailed results and discussion pertaining to stimulus validation can be found in Supplemental Materials: Stimulus Creation Process. The stimuli were first rated by two independent raters for realism on a 1 to 9 Likert scale to narrow down the stimuli to 6 per condition. After this, to further examine the stimuli, a pilot version of our previous work (Lawrence, Cimermanis, & Collin, 2025) was run. Using the by-stimulus data, we informally selected the top 3 AIFs from each condition which were most often misidentified as real. This ensured that our final set of 72 AI-generated images were the most realistic ones from the initial generation. In two cases, we did not select the third highest rated face because the angry emotional expression was very rarely misidentified as real. As such, we opted to keep the fourth highest rated face, which was overall slightly lower, but was more consistently rated across emotional expressions.

The photographs of real faces (RFs) were drawn from the RADIATE Database (Conley et al., 2018; Tottenham et al., 2009), which contains images of men and women who self-identified as Asian, Black, Hispanic, and White. The individuals express eight emotions (angry, calm, disgusted, fearful, happy, neutral, sad, and surprised), each with closed-mouth and open-mouth variations. Half the images are of women, and half of men. The images used in the experiment were showing four emotional states: angry, fearful, happy, and sad. The individuals were Asian, Black, or White. All facial expressions were the open-mouthed variants. All images were resized to be 375 × 375px. We administered a demographics questionnaire of our own design, and the Prosopagnosia Index (Shah et al., 2015).

### *3.3.3. Procedure*

The study was conducted online using Qualtrics™ and involved participants making judgments of AIFs and RFs. Participants were not informed that the images they saw were a combination of RFs and AIFs. Indeed, instructions made no mention of AI-generated faces, though participants were fully debriefed following the experiment.

Participants first completed the demographics questionnaire and the Prosopagnosia Index (Shah et al., 2015). Next, they were given instructions on the task. Stimuli were displayed in random order. Each image was presented with six rating scales below it, where participants indicated the extent to which they felt that the face was expressing several emotions. The emotions they rated were angry, happy, disgusted, fearful, sad, and surprised (See Figure 2 for a trial example). These emotions included the four target emotions represented in the stimuli, as well as two filler emotions, disgusted and surprised, as they are often confused with anger and fear, respectively (Y. Wang et al., 2019). Participants rated each emotion on a scale from 0 (emotion absent) to 5 (very strong emotion). Neutral was not included as a rating scale; instead, a score of 0 on all rating scales considered neutral, and participants were informed to set all scales to zero to give a rating of neutral. Each image was shown on screen until participants submitted their responses, followed by a 1 second inter-stimulus interval. Images were presented at a size of 375 by 375 pixels. On the computer in our lab, these subtended 10.6 by 10.6 degrees, but this would have varied somewhat across participants, who completed the task on a computer of their choosing.

After rating all the stimuli, participants were asked whether they were distracted during the experiment. Next, they were debriefed on the true manipulation of the study, and they had the option to have their data excluded from analyses. Finally, participants were asked about their familiarity with AI software and were able to provide general comments on the study.

**Figure 2**

*Example Trial Screen for Study 2*

| | Absent Emotion 0 | 1 | 2 | 3 | 4 | Very Strong Emotion 5 |
|---|---|---|---|---|---|---|
| Angry | ○ | ○ | ○ | ○ | ○ | ○ |
| Disgusted | ○ | ○ | ○ | ○ | ○ | ○ |
| Happy | ○ | ○ | ○ | ○ | ○ | ○ |
| Fearful | ○ | ○ | ○ | ○ | ○ | ○ |
| Sad | ○ | ○ | ○ | ○ | ○ | ○ |
| Surprised | ○ | ○ | ○ | ○ | ○ | ○ |

*Note.* Participants viewed a 375px-by-375px image of a face, below it were rating scales for the 6 basic emotions. Participants rated each emotion from 0 (emotion absent) to 5 (very strong emotion), to give a 'neutral' rating, participants were instructed to set all scales to 0.

### *3.3.4. Data Analysis*

Three dependent variables were extracted from the raw data: Accuracy, Intensity, and Saliency. These DVs were calculated as described below. There were slight deviations from the

pre-registration in terms of data cleaning based on catch trials[4] and outlier rejection (see below for details). Data has been made available here: https://osf.io/e4m2y/overview?view_only=2e5fc829bcd94c12a1cc3a605f445083.

A participant's ratings were considered accurate if they gave a higher rating to the target emotion than to any other. For instance, for an angry face, if the rating of Anger was higher than the rating of any other emotion, that trial was scored as a 1 for accuracy, otherwise it was scored as 0. Accuracy was calculated after first replacing any missing values with a value of zero. The accuracy scores were then averaged within each condition.

Intensity was calculated by taking the average of the target emotion's ratings within each condition, ignoring the ratings given to other emotions. For instance, for a happy face, if the participant gave a Happy rating of 4, then that was considered the intensity for that condition, regardless of what ratings were given on the other 5 emotion scales.

Saliency was calculated by taking the ratio of the target emotion rating to the sum of all ratings. For instance, for a Fearful face, if the participant gave ratings of Angry = 1, Disgusted = 0, Happy = 0, Fearful = 4, Sad = 1, and Surprised = 1, then the Saliency would be equal to 4/(1+0+0+4+1+1) = .57. These ratios were then averaged within each condition.

Before any analysis occurred, the data was first cleaned using several criteria. First, participants were removed based on outlying total task duration +/-2.5SD from the mean (n = 15 removed). At this point, participants would have been removed if they answered less than ¾

[4] Catch trials refers to 4 trials where the stimulus had a number overlaid on top of it. Participants were informed these trials would appear randomly throughout the experiment. When they saw these trials, they were instructed to set every scale to the number which was overlaid on the stimulus, regardless of what their actual rating would have been.

catch trials correctly. We opted not to use this criterion as this resulted in over 50% of participants being removed. Moreover, during recruitment, we noted participant confusion when the catch trials appeared. Based on this, it appears the catch trials did not function as intended, and we would be removing participant data that was otherwise sound. We therefore opted to deviate from the pre-registration and not remove participants based on catch trials. For outlier Winsorizing/cleaning, each DV was handled separately. We deviated from the pre-registration by opting to remove participants who were marked as outliers on one DV from all DVs. This was done to increase the stringency of data rejection, since participants were no longer being removed based on catch trials. All outlier analyses were run using the Van Selst and Jolicoeur (1994) guidelines. For accuracy, an overall accuracy score was calculated for each participant. If their score fell outside of the cutoff, that participant was removed (n = 24). For saliency, participants were removed if 5 or more saliency scores fell outside of the cutoff (n = 0). If a participant had 4 or fewer outlying scores, those scores were Winzorized. Finally, for intensity, participants were removed if 5 or more intensity scores fell outside of the cutoff (n = 0). If a participant had 4 or fewer outlying scores, those scores were Winzorized. This resulted in our final sample size of N = 450.

Following outlier analysis, a 2 × 2 × 3 × 4 repeated measures ANOVA was run with Image Type (RF or AIF), Stimulus Gender (man or woman), Stimulus Race (Asian, Black, or White), and Stimulus Emotional Expression (anger, fear, happiness, or sadness) as independent variables. Separate analyses were run for each DV, which, as stated above, were accuracy, intensity, and saliency. These dependent variables have been drawn from previous literature examining faces and emojis (Beaupré & Hess, 2005; Boutet et al., 2023).

### 3.4. Results

Results were analyzed via a series of ANOVAs, starting with an omnibus ANOVA for each of the three DVs. All main effects and interactions in the omnibus ANOVA were significant at $p < .001$. As such, we mainly interpreted the results by visual inspection of the graphs. However, a series of ANOVAs was conducted to reinforce these interpretations. Details of the ANOVAs and post-hoc tests can be found in the supplementary materials tables S2-S23. For all analyses, where the assumption of sphericity was violated, Greenhouse-Geisser adjusted values are reported.

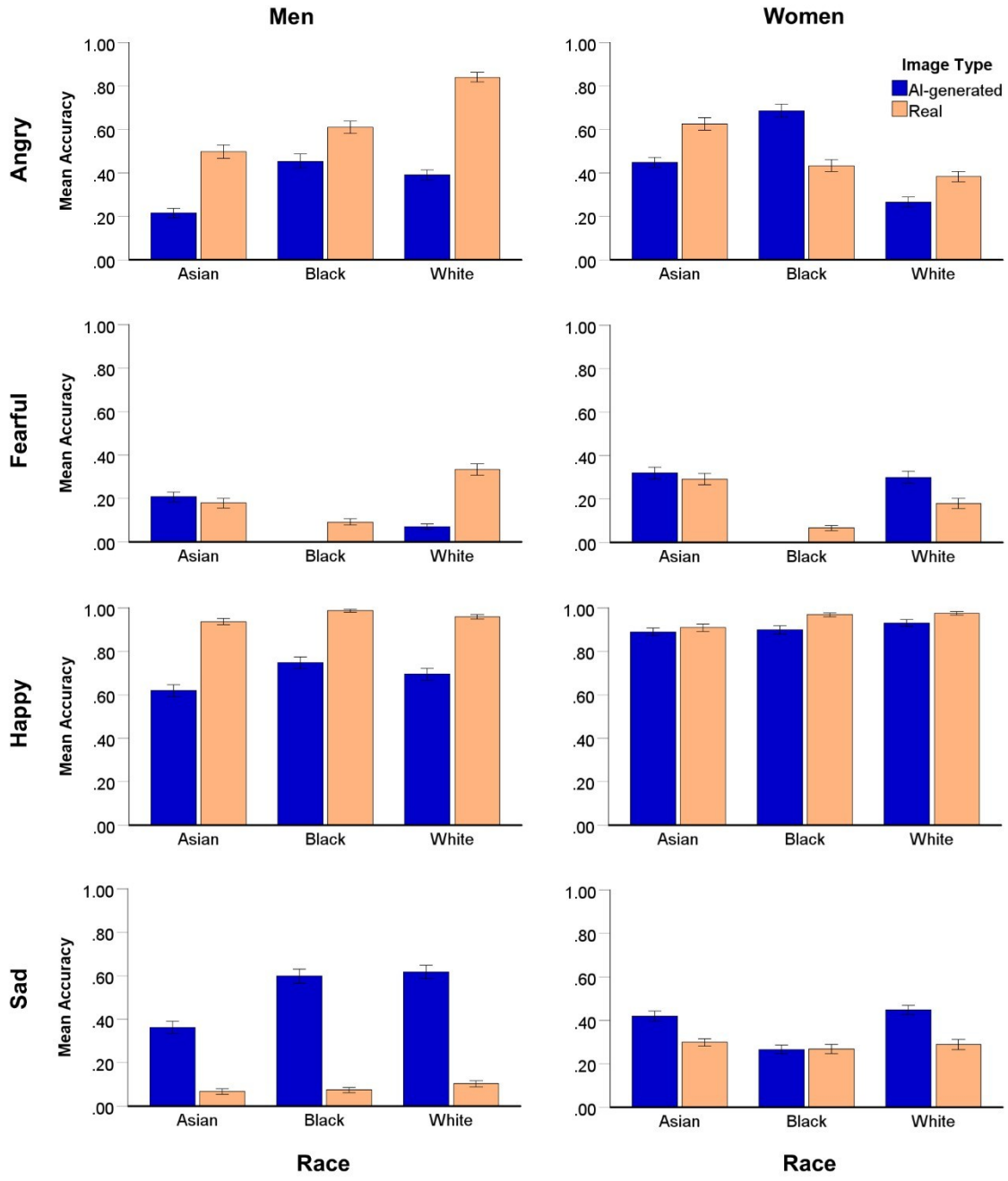
### *3.4.1. Accuracy*

Mean accuracy results (with 95% confidence intervals) are illustrated in Figure 3. As can be seen there, RFs generally yielded higher accuracy values than AIFs for expressions of Happiness and Anger. This effect for happiness was much larger for men's faces than women's. The effect of image type in the Anger condition was reversed in Black women's faces, such that the AI-generated ones yielded higher accuracy than real faces. The effect of image type was also reversed for sad faces. That is, AI-generated sad faces were more readily seen as sad than were real ones. This effect was much larger for men's faces compared to women's faces generally, though there was no difference between RFs and AIFs for Black women's sad faces.

Generally, for men's faces, accuracy differences between races followed the same pattern, where emotions of Asian faces were least accurately recognized, followed by Black faces, with the highest accuracy being for White faces. For women's faces, no such pattern emerged between races.

**Figure 3**

*Mean Accuracy of Target Emotion Identification for AI-generated vs. Real Faces*




*Note.* Mean accuracy for target emotions. Seven emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) were rated on 6-point Likert scales ranging from 0 (Absent Emotion) to 5 (Very Strong Emotion). A rating of 0 on all scales was considered neutral. Error bars represent 95% CI.

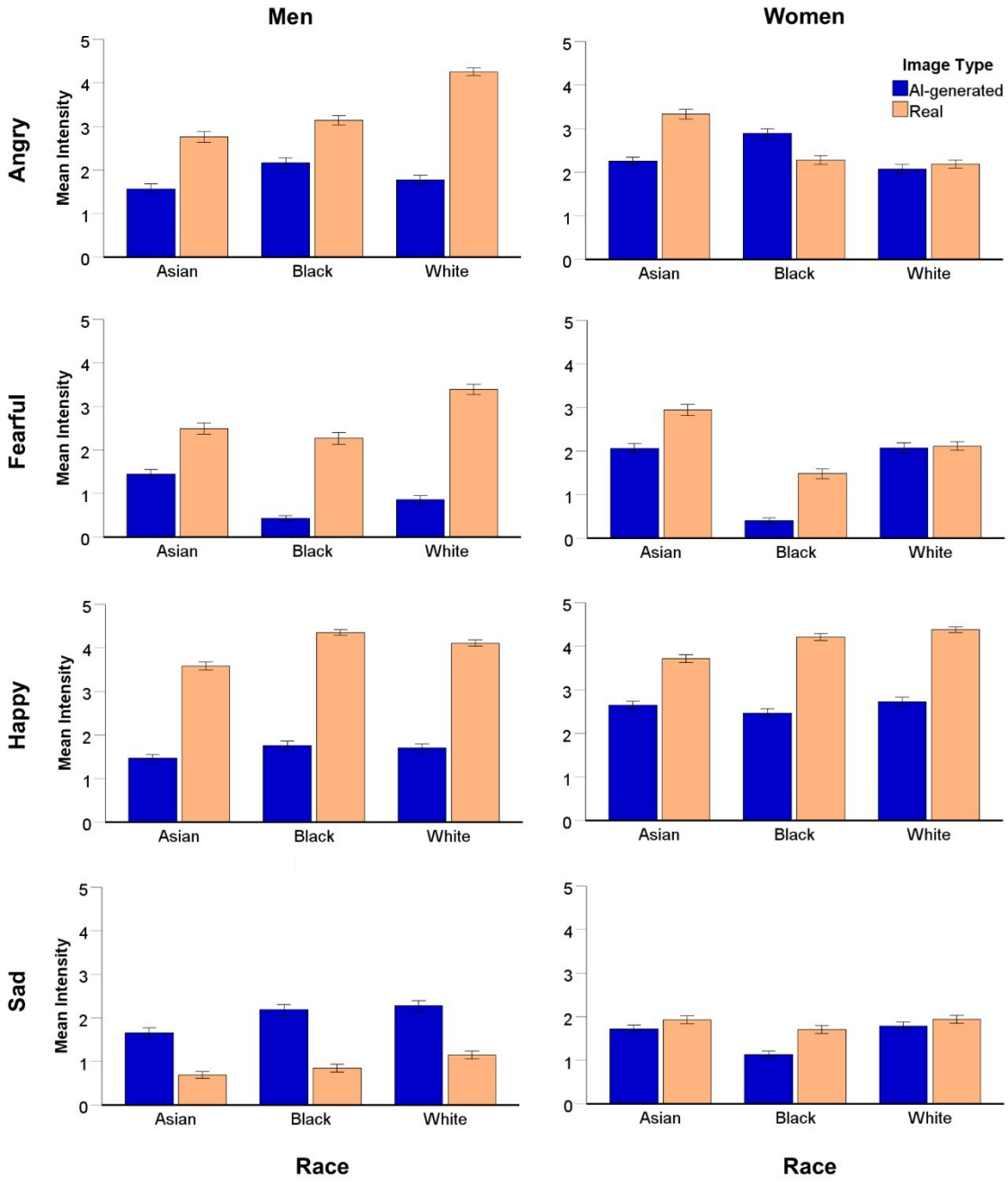
### *3.4.2. Intensity*

Mean intensity results with 95% confidence intervals are illustrated in Figure 4. As can be seen there, RFs generally yielded higher intensity ratings compared to AIFs for all emotions, races, and genders, aside from sadness ratings of men's faces, and anger ratings of Black women's faces, which were in the opposite direction, and fear for White women's faces, in which there was no difference between image types. Again, the effect for happy and angry faces is larger for men compared to women.

Generally, for men's faces, intensity between races followed the same pattern, where emotions of Asian faces were least intense, followed by Black faces, with the highest intensity being for White faces. For women's faces, no such pattern emerged between races.

**Figure 4**

*Mean Intensity of Target Emotions for AI-generated vs. Real Faces*




*Note.* Mean intensity for target emotions. Seven emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) were rated on 6-point Likert scales ranging from 0 (Absent Emotion) to 5 (Very Strong Emotion). A rating of 0 on all scales was considered neutral. Error bars represent 95% CI.

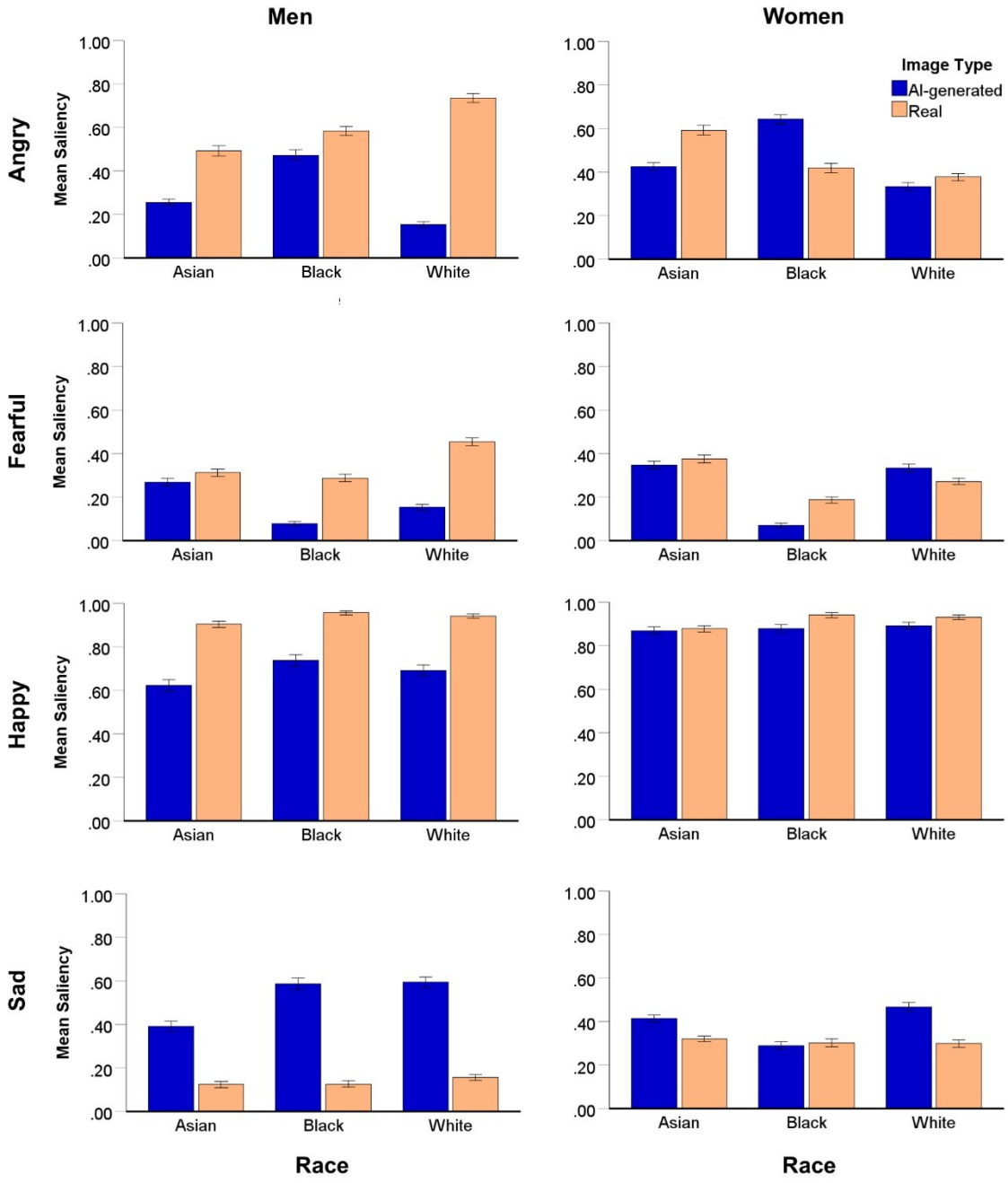
### *3.4.3. Saliency*

Mean saliency results are illustrated in Figure 5. As can be seen there, RFs generally yielded higher saliency ratings compared to AIFs for emotions, races, and genders, aside from sadness ratings of men's faces in general and Asian and White women's faces, anger ratings of Black women's faces, and fear for White women's faces, which were in the opposite direction. Again, for angry and happy faces, the effect was larger for men's compared to women's faces. Moreover, there is no longer a difference between image types for Asian women's happy faces.

Generally, for men's faces, saliency between races followed the same pattern where target emotions of Asian faces were least salient, followed by Black faces, with the highest saliency being for White faces. For women's faces, no such pattern emerged between races.

**Figure 5**

*Mean Saliency of Target Emotions for AI-generated vs. Real Faces*




*Note.* Mean saliency for target emotions. Seven emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) were rated on 6-point Likert scales ranging from 0 (Absent Emotion) to 5 (Very Strong Emotion). A rating of 0 on all scales was considered neutral. Error bars represent 95% CI.

### 3.5. Discussion

The goal of the present study was to examine the biases in an AI-generator's ability to generate convincing emotional expressions across different races and genders. Our hypotheses and research questions are restated hereunder. While discussing the hypotheses, we also explored the possible stimulus race and gender effects.

***H1:*** *Emotion recognition will be poorer with AI-generated faces than real faces, particularly for the AI-generated faces of people of colour compared to the White AIFs.*

The results show mixed support for H1. Emotion recognition was not poorer for AIFs compared to RFs across all conditions.

For angry expressions, those of AIFs were less accurately recognized than those of real faces in all gender-by-race conditions, aside from Black women. We speculate that this could point to the AI-generator training being biased towards the stereotype of the "angry black woman" which, despite a lack of supporting empirical evidence, is widely perpetuated in media (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009).

For expressions of fear, those of AIFs were recognized equally well or less accurately than those of real faces in all conditions, aside from White women. This difference could be due to a larger proportion of fear expressions being portrayed on faces of White women in the media that the AI-generators are trained on. Horror movies often portray White women as the "final girl" or central character (Quigley, 2023).

Accuracy in identifying expressions of happiness was equal or poorer for AIFs as compared to real faces. This was true across all conditions.

For sad faces, the expressions of AIFs were recognized more accurately than those of real faces in all conditions aside from Black women. Within men's faces, the difference is much smaller for Asian males. This could be due, perhaps, to AI-generators being trained on datasets that include more intense and genuine expressions of sadness from men (e.g., news media).

Interestingly, differences in accuracy tended to be much larger for men's faces compared to women's faces. This appears to be due to men's AIFs having lower scores compared to women's AIFs, rather than differences between the RFs of men and women. In contrast, this difference seems to come from a greater accuracy in identifying emotions of AIFs for men compared to women when emotions of AIFs were recognized more accurately than RFs.

Overall, we see a trend of higher accuracy ratings for White faces compared to Black and Asian faces. This is likely due to the large representation of White faces in training sets of AI-generators (Kärkkäinen & Joo, 2021).

***H2:*** *Emotion intensity ratings will be lower for AI-generated faces than real faces, particularly for the AI-generated faces of people of colour compared to the White AIFs.*

H2 was partially supported. Similar trends were found for emotional intensity as were found for accuracy. That is, for anger, AI-generated faces were less intensely or equally intensely rated as compared to real faces, aside from the Black women's faces. Again, this could be due to the prominent, albeit unfounded, stereotype of "the angry Black woman" (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009), which the AI-generator may be trained on, thus creating a bias in generating images of angry Black women.

For fearful faces, the differences in intensity are much more prominent compared to those for accuracy. In all conditions, aside from White women, AIFs were rated less intensely than

RFs. Again, this suggests that there may be a larger proportion of fearful White women in the training sets. One possible explanation for this, albeit a speculative one, is that the data sets contain a large number of images from films, including horror films, where a screaming white woman is a trope of western cinema.

For happiness, AIFs again were rated less intensely compared to RFs. This could be due to expressions of happiness in media which the generator is trained on being largely posed. If this is the case, then the resulting images would then also tend to appear posed, which may make them appear less intense.

For sadness, for men's faces, AIFs were rated *more* intensely than the RFs, whereas for women, there was either no difference, or AIFs were less intensely rated than RFs. As with accuracy, this could be due to a large proportion of publicly available images of intense and genuine sadness (e.g., news stories which depict people after experiencing, witnessing, or learning of a tragedy). This may also be due to social display rules in which men are not expected to display emotions of sadness (Kret & De Gelder, 2012), so the real photos may not depict sadness as accurately as the women's photos.

Interestingly, in all cases, the effect of AIFs vs. RFs tends to be larger for men's faces. This difference seems to arise due not only to AIFs being perceived as less intense than AIFs of women, but also that RFs of men are perceived as more intense than those of women (aside from sadness).

Moreover, we see a trend of higher intensity ratings for White faces compared to Black and Asian faces. This is likely due to the large representation of White faces in training sets of AI-generators.

***H3:*** *Saliency measures will be lower for AI-generated faces than real faces, particularly for the AI-generated faces of people of colour compared to the White AIFs.*

H3 was partially supported. Again, similar trends to the other dependent variables were found. For anger, AIFs had lower saliency compared to RFs, aside from Black women's faces. Again, this could be due to the prominent, albeit unfounded, stereotype of "the angry Black woman" (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009), which the AI-generator may be trained on, thus creating a bias in generating images of angry Black women.

For fear, in all conditions, aside from White women, AIFs were less salient than RFs. Again, this may be due to a larger proportion of fearful White women in the training sets (e.g., leading actors in horror films).

For happiness, AIFs again were less salient (or for women's faces equally as salient) compared to RFs. Although they were less salient, AIFs overall had quite high levels of saliency. This could be due to the large dataset which generators are trained on, where some images could have multiple emotion labels, thus creating noise in the training set which impacts the resulting images and their emotional saliency (Kärkkäinen & Joo, 2021).

For sadness, in all conditions, aside from Black women's faces, AIFs are more salient than RFs. Again, this could be due to the intense and genuine expressions of sadness that AI-generators can be trained on. It may also arise from sadness being a less natural emotion to pose, thus the RFs may have been less representative of sadness as an isolated emotion.

Again, in most conditions, the difference for men was much more pronounced than for women, especially for happiness where there was almost no difference between saliency of AIFs

and RFs. This is in contrast to previous work which has found that generated images of men tend to be higher quality than those of women, since the training sets include a large proportion of men, particularly White men (AlDahoul et al., 2025; Bianchi et al., 2023; Ghosh & Caliskan, 2023; Kärkkäinen & Joo, 2021; Nightingale & Farid, 2022).

Moreover, we see a trend of higher saliency ratings for White faces compared to Black and Asian faces. This is likely due to the large representation of White faces in training sets of AI-generators.

***H4:*** *Happy faces will yield the best results compared to the other emotional expressions such that they will yield better emotion recognition, higher intensity, and higher saliency.*

Recognition, intensity, and saliency for happiness, despite generally being lower for AIFs, was still quite high. By far it was the most accurately recognized and most highly rated in terms of intensity and saliency. Moreover, recognition, intensity, and saliency of happiness in AIFs for men's faces was lower compared to women's faces. This may indicate that smiling men's faces are less represented in training sets of AI-generators compared to smiling women, which contrasts previous findings (AlDahoul et al., 2025; Bianchi et al., 2023; Ghosh & Caliskan, 2023; Kärkkäinen & Joo, 2021; Nightingale & Farid, 2022). Moreover, the difference between races is not as prominent for happy faces. When there is a difference, often it is a reduction in accuracy, intensity, and saliency for Asian faces. However, to reiterate from above, although larger training sets are seen to increase realism of photos, it is possible that the larger datasets introduce a certain level of noise when it comes to the emotional expression, and the AI-generators are favouring realism over intense expressions.

### 3.6. Limitations

One limitation of this study was using images which were not FACS (facial action; Ekman & Friesen, 1978) coded. FACS coding ensures that images are in fact displaying the emotions that they intend to. While the RADIATE dataset (Conley et al., 2018; Tottenham et al., 2009) was validated, FACS coding would provide a more objective measure of the emotions. Future studies should use FACS coded images when examining the differences between AIFs and RFs.

A second limitation was using a convenience sample of undergraduate students. Because of this, the large proportion of female and White students prevented us from doing more analyses based on participant gender and race. Future studies should recruit more diverse samples in order to test the generalizability of the findings.

A final limitation was that the attention checks employed did not work as intended. As such, we chose not to exclude participants based on this criterion, deviating from the pre-registration. Future studies, especially those conducted online, should use attention checks to improve data quality.

### 3.7. Conclusion

In conclusion, when using the same simple prompts to generate images of faces of varying races, genders, and emotional expressions, there are differences in the quality of the images. In some cases, we also find an advantage of men over women, however this finding was not consistent across conditions. These findings indicate that the training sets of AI-generators are in fact biased, with large proportions of White faces. Whether there is a bias towards generating men or women's faces has not been supported in either direction in our study. Biases when generating images of varying races and genders may have real-world implications as AI-generated images become more commonplace. If we are regularly viewing images of different

groups in stereotyped or biased representations, that may impact how we start to think about and view those groups in real life. This study provides further evidence that AI-generators are replicating biased views of certain groups.

**3.8. References**

AlDahoul, N., Rahwan, T., Zaki, Y. (2025) AI-generated faces influence gender stereotypes and racial homogenization. Sci Rep 15(1):14449. https://doi.org/10.1038/s41598-025-99623-3

American Psychological Association, APA Task Force on Race and Ethnicity Guidelines in Psychology. (2019). Race and Ethnicity Guidelines in Psychology: Promoting Responsiveness and Equity. Retrieved from http://www.apa.org/about/policy/race-and-ethnicity-in-psychology.pdf

Ashley, W. (2014). The angry black woman: The impact of pejorative stereotypes on psychotherapy with black women. *Social work in public health, 29*(1), 27-34. DOI: 10.1080/19371918.2011.619449

Assis, J. D., & Moura, M. A. (2025). Algorithmic Semiosis and Racial Bias: A Study of Images Created by Generative AI. *Encontros Bibli*, *30*, e103495. DOI: 10.5007/1518-2924.2025.e103495

Beaupré, M. G., & Hess, U. (2005). Cross-Cultural Emotion Recognition among Canadian Ethnic Groups. *Journal of Cross-Cultural Psychology*, *36*(3), 355–370. https://doi.org/10.1177/0022022104273656

Bhardwaj, N., Bhardwaj, A., & Garg, L. (2025, February). Controlling Bias in Generative AI: Techniques for Fair and Equitable Data Generation in Socially Sensitive Applications. In *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1-8). IEEE

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation

Amplifies Demographic Stereotypes at Large Scale. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. https://doi.org/10.1145/3593013.3594095

Boutet, I., Guay, J., Chamberland, J., Cousineau, D., & Collin, C. (2023). Emojis that work! Incorporating visual cues from facial expressions in emojis can reduce ambiguous interpretations. *Computers in Human Behavior Reports*, *9*, 100251. https://doi.org/10.1016/j.chbr.2022.100251

Campbell, D. W., Neuert, T., Friesen, K. B., & McKeen, N. A. (2010). Assessing Social Approachability: Individual Differences, In-Group Biases, and Experimental Control. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *42*(4), 254–263. https://doi.org/10.1037/a0020229

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, *270*, 1059–1067. https://doi.org/10.1016/j.psychres.2018.04.066

Eiserbeck, A., Maier, M., Baum, J., & Abdel Rahman, R. (2023). Deepfake smiles matter less—The psychological and neural impact of presumed AI-generated faces. *Scientific Reports*, *13*(1), 16111. https://doi.org/10.1038/s41598-023-42802-x

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*. https://psycnet.apa.org/doiLanding?doi=10.1037/t27734-000.

Gawali, A., Tarle, A., Can Inac, C., & Kulkarni, S. S. (2024). A Framework for Detecting Stereotypes, Prejudices and Discrimination in AI-Generated Imagery.

Ghosh, S., & Caliskan, A. (2023). 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *Findings of the Association for*

*Computational Linguistics: EMNLP 2023*, 6971–6985. https://doi.org/10.18653/v1/2023.findings-emnlp.465

Gorska, A. M., & Jemielniak, D. (2023). The invisible women: uncovering gender bias in AI-generated images of professionals. *Feminist Media Studies*, *23*(8), 4370-4375. https://doi.org/10.1080/14680777.2023.2263659

Jääskeläinen, P., Sharma, N. K., Pallett, H., & Åsberg, C. (2025). Intersectional analysis of visual generative AI: the case of stable diffusion. *AI & SOCIETY*, 1-22. https://doi.org/10.1007/s00146-025-02207-y

Jiang, Z., Recio, G., Li, W., Zhu, P., He, J., & Sommer, W. (2023). The other-race effect in facial expression processing: Behavioral and ERP evidence from a balanced cross-cultural study in women. *International Journal of Psychophysiology*, *183*, 53–60. https://doi.org/10.1016/j.ijpsycho.2022.11.009

Kärkkäinen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. Proceedings/IEEE Workshop on Applications of Computer Vision, 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159

Kret, M. E., & De Gelder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia*, *50*(7), 1211–1221. https://doi.org/10.1016/j.neuropsychologia.2011.12.022

Lawrence, M., Cimermanis, K. N. E., & Collin, C. A. (2025). Not all AI-generated faces are created equal: impacts of model gender, race, and emotional expression on classification accuracy. *AI & SOC*, 1-12. https://doi.org/10.1007/s00146-025-02670-7

Lee, K. (2013). Why Asian female stereotypes matter to all: Beyond black and white, east and west. *Critical Philosophy of Race, 1*(1), 86-103.

Liefooghe, B., Oliveira, M., Leisten, L. M., Hoogers, E., Aarts, H., & Hortensius, R. (2023). Are Natural Faces Merely Labelled as Artificial Trusted Less? *Collabra: Psychology*, *9*(1), 73066.

Locke, L. G., & Hodgdon, G. (2024). Gender bias in visual generative artificial intelligence systems and the socialization of AI. *AI & SOCIETY*, 1-8. https://doi.org/10.1007/s00146-024-02129-1

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin*, *126*(3), 424–453. https://doi.org/10.1037/0033-2909.126.3.424

Melson-Silimon, A., Spivey, B. N., & Skinner-Dorkenoo, A. L. (2024). The construction of racial stereotypes and how they serve as racial propaganda. *Social and Personality Psychology Compass, 18*(1), e12862. DOI: 10.1111/spc3.12862

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. https://doi.org/10.1007/s10339-005-0050-6

Morgan, M., & Bennett, D. (2006). Getting off of Black women's backs: Love her or leave her alone. *Du Bois Review: Social Science Research on Race, 3*(2), 485-502. https://doi.org/10.1017/S1742058X06060334

Mubashir, M. (2024). The gendered dress of DALL-E 2: Exploring profession-based images in the Indian context. *MedieKultur: Journal of media and communication research*, *40*(76), 100-119. https://doi.org/10.7146/mk.v40i76.143565

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

Patil, S., Cuenca, P., Lambert, N., & von Platen, P. (2022, August 22). *Stable Diffusion with 🧨 Diffusers*. Hugging Face. https://huggingface.co/blog/stable_diffusion

Park, Y. S. (2024). White default: Examining racialized biases behind AI-generated images. *Art Education*, *77*(4), 36-45. https://doi.org/10.1080/00043125.2024.2330340

Quigley, J. (2023). *The Call from Inside the House: The Final Girl Trope as a Reflection of Cultural Anxiety of Race and Gender* (Master's thesis, Idaho State University).

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343. https://doi.org/10.1098/rsos.140343

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

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

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science*, *18*(3), 184–188.

Walley-Jean, J. C. (2009). Debunking the myth of the “angry Black woman”: An exploration of anger in young African American women. *Black Women, Gender & Families*, *3*(2), 68–86.

Wang, Y., Zhu, Z., Chen, B., & Fang, F. (2019). Perceptual learning and recognition confusion reveal the underlying relationships among the six basic emotions. *Cognition and Emotion*, *33*(4), 754–767. https://doi.org/10.1080/02699931.2018.1491831

Wu, Y., Nakashima, Y., & Garcia, N. (2025). Revealing gender bias from prompt to image in stable diffusion. *Journal of Imaging*, *11*(2), 35. doi: 10.3390/jimaging11020035

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2011). Perception and motivation in face recognition: A critical review of theories of the Cross-Race Effect. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *16*(2), 116–142. https://doi.org/10.1177/1088868311418987

## 3.9. Supplementary Materials

### Stimulus Creation Process

Stimuli were created for this study using Leonardo.ai, which is a website that implements Stable Diffusion (Patil et al., 2022). Because of this, stimuli were validated by showing participants a set of real faces mixed in with the AI-generated images and instructing them to select whether the image on screen was AI-generated or real. N = 89 participants were recruited from the University of Ottawa's Integrated System for Participation in Research (ISPR). First, data was organized and cleaned using several criteria. First, participants were removed if they incorrectly answered 2 or more catch trials (n = 23). Next, participants were removed if they had outlying durations (n = 2; using the VanSelst and Jolicoeur, 1994 guidelines). Finally, accuracy scores were calculated for each condition (i.e., 48 accuracy scores for each participant). Participants were removed if 5 or more accuracy scores were outlying (n = 3; using VanSelst and Jolicoeur guidelines). This resulted in a final sample size of N = 61.

After data cleaning, several sums were calculated. First, for each stimulus, one sum was calculated which showed the total number of participants that misidentified an AI-generated faces as real. Second, a sum was calculated for each stimulus model by summing the total number of misidentified AI-generated faces across the four emotions for each model. For example, one set of images was generated from the RADIATE database (Conley et al., 2019; Tottenham et al., 2008) of model number 6 who self-identified as an Asian woman. Therefore, to get the total sum for this model, the AI-generated faces for anger, fear, happiness, and sadness were added together. The higher the total sum, the more often the model was misidentified as real, rather than correctly identified as AI-generated. Therefore, the three models with the highest sums in each condition were kept for the main study. In two cases, the third highest sums were

not kept, and instead replaced by the models with the fourth highest in their respective conditions, due to the angry emotional expression being very rarely misidentified as real. As this was the case, we opted to keep the fourth highest model whose total misidentification was lower, but more consistent across the four emotional expressions.

As a final step, we returned to the data before any cleaning had occurred and d' values were calculated using only the selected stimuli (N = 89). Following this, data was cleaned following a similar procedure as above. First, participants were removed if they answer 2 or more catch trials incorrectly (n = 23). Second, participants were removed if they had an outlying duration (n = 2). Finally, participants' d' values were winsorized if they had 3 or less outlying scores, and the participant was removed entirely if they had 4 or more outlying scores (n = 2). This resulted in a final sample size of N = 62. Using this data, one-sample t-tests were run on each condition, comparing the mean to a value of 0 (chance level) to assess whether the AI-generated faces were being detected above chance level. All comparisons were statistically significant ($p < .001$, Cohen's d ranging from .48 to 2.00). While this indicates that the faces were detected above chance level, this could be because participants knew to look for them. In the current study, participants were not told that AI-generated faces were present.

**Table S1**

*Demographic Breakdown of Participants*

| | Mean | SD |
|---|---|---|
| Age (Years) | 20.01 | 3.65 |
| Education (Years) | 1.89 | 1.37 |
| Time spent in Canada (Years) | 8.24 | 6.30 |
| | N | Percentage (%) |
| Race | | |
| Arab | 35 | 7.78 |
| Arab/African | 13 | 2.89 |
| Asian-East | 53 | 11.78 |
| Asian-South | 62 | 13.78 |
| Asian-Southeast | 18 | 4.00 |
| Black-African | 38 | 8.44 |
| Black-Caribbean | 12 | 2.67 |
| Black-North American | 4 | .89 |
| First Nation | 5 | 1.11 |
| Indian-Caribbean | 1 | .22 |
| Indigenous/Aboriginal | 2 | .44 |
| Inuit | 0 | 0 |
| Latin American | 15 | 3.33 |
| Metis | 3 | .67 |
| Middle Eastern | 30 | 6.67 |
| Mixed Heritage | 22 | 4.89 |
| Pacific Islander | 0 | 0 |
| White-North American | 142 | 31.56 |
| White-European | 77 | 17.11 |

| | | |
|---|---|---|
| I don't know | 1 | .22 |
| You don't have an option that applies to me | 4 | .89 |
| Choose not to respond | 3 | .67 |
| Gender | | |
| Agender | 8 | 1.78 |
| Cisgender Female | 331 | 73.56 |
| Cisgender Male | 90 | 20.00 |
| Genderfluid | 2 | .44 |
| Genderqueer | 3 | .67 |
| Non-binary | 5 | 1.11 |
| Questioning | 2 | .44 |
| Transgender Female | 0 | 0 |
| Transgender Male | 2 | .44 |
| Two Spirit | 0 | 0 |
| No Option Applies | 6 | 1.33 |
| Choose not to respond | 9 | 2.00 |

*Note.* Mean and Standard Deviation (SD) for age, years of education (n = 447), and time spent in Canada (n = 156) for participants. For time spent in Canada, only those who reported living somewhere else were included. Number of participants (N) and percentage of race and gender of participants (N = 450). The total percentage for race and gender may not equal 100% as participants were instructed to select all that apply, so some participants selected more than one option, or due to missing data. Although Mixed Heritage was an option, not all participants who selected multiple options selected this, and not all participants who selected Mixed Heritage specified further.

**Table S2**

*Accuracy: Results for the 2 × 2 × 3 × 4 Repeated Measures ANOVA*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 29.42$ | < .001 | .06 |
| Simulus Gender | $F(1,449) = 168.48$ | < .001 | .27 |
| Stimulus Race | $F(2,898) = 81.50$ | < .001 | .15 |
| Emotional Expression | $F(2.63,1179.91) = 3326.04$ | < .001 | .88 |
| Image Type*Stimulus Gender | $F(1,449) = 184.63$ | < .001 | .29 |
| Image Type*Stimulus Race | $F(1,898) = 57.42$ | < .001 | .11 |
| Image Type*Emotional Expression | $F(2.75,1235.30) = 569.01$ | < .001 | .59 |
| Stimulus Gender*Stimulus Race | $F(2,898) = 409.05$ | < .001 | .48 |
| Stimulus Gender*Emotional Expression | $F(2.81,1263.38) =$ | < .001 | .18 |
| Stimulus Race*Emotional Expression | $F(5.59,2510.81) = 240.46$ | < .001 | .35 |
| Image Type *Stimulus Gender*Stimulus Race | $F(2,898) = 56.74$ | < .001 | .11 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.70,1210.39) =$ | < .001 | .57 |
| Image Type*Stimulus Race*Emotional Expression | $F(5.43,2439.13) = 130.61$ | < .001 | .23 |
| Stimulus Gender*Stimulus Race*Emotional Expression | $F(5.48,2458.75) = 176.77$ | < .001 | .28 |
| Image Type*Stimulus Gender*Stimulus Race*Emotional Expression | $F(5.59,2510.55) = 94.81$ | < .001 | .17 |

*Note.* N = 450.

**Table S3**

*Accuracy: Results for the 2 × 2 × 4 Repeated Measures ANOVA within Asian Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 51.11$ | < .001 | .10 |
| Simulus Gender | $F(1,449) = 871.18$ | < .001 | .66 |
| Emotional Expression | $F(2.83,1268.31) = 1465.43$ | < .001 | .77 |
| Image Type*Stimulus Gender | $F(1, 449) = 40.81$ | < .001 | .08 |
| Image Type*Emotional Expression | $F(2.77,1245.56) = 258.62$ | < .001 | .37 |
| Stimulus Gender*Emotional Expression | $F(2.95,1323.54) = 10.65$ | < .001 | .02 |
| Image Type*Stimulus Gender*Emotional Expression | $F(3,1347) = 106.85$ | < .001 | .19 |

*Note.* N = 450.

**Table S4**

*Accuracy: Results for the 2 × 2 × 4 Repeated Measures ANOVA within Black Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 12.27$ | < .001 | .03 |
| Simulus Gender | $F(1,449) = .50$ | .48 | .001 |
| Emotional Expression | $F(2.26,1014.94) = 3216.16$ | < .001 | ,88 |
| Image Type*Stimulus Gender | $F(1,449) = 5.09$ | .025 | .01 |
| Image Type*Emotional Expression | $F(2.36,1060.96) = 295.84$ | < .001 | .40 |
| Stimulus Gender*Emotional Expression | $F(2.23,1000.34) = 41.46$ | < .001 | .09 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.32,1039.30) = 491.51$ | < .001 | .52 |

*Note.* N = 450.

**Table S5**

*Accuracy: Results for the 2 × 2 × 4 Repeated Measures ANOVA within White Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 63.38$ | < .001 | .12 |
| Simulus Gender | $F(1,449) = 40.80$ | < .001 | .08 |
| Emotional Expression | $F(2.79,1252.43) = 2136.85$ | < .001 | .83 |
| Image Type*Stimulus Gender | $F(1, 449) = 258.70$ | < .001 | .37 |
| Image Type*Emotional Expression | $F(2.81,1262.09) = 575.41$ | < .001 | .56 |
| Stimulus Gender*Emotional Expression | $F(2.90,1300.74) = 385.69$ | < .001 | .46 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.83,1269.28) = 316.87$ | < .001 | .41 |

*Note.* N = 450.

**Table S6**

*Accuracy: Results for the 2 × 4 Repeated Measures ANOVA within Asian Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 89.89$ | < .001 | .17 |
| Emotional Expression | $F(2.87,1287.69) = 1009.71$ | < .001 | .69 |
| Image Type*Emotional Expression | $F(2.92,1310.36) = 314.03$ | < .001 | .41 |

*Note.* N = 450.

**Table S7**

*Accuracy: Results for the 2 × 4 Repeated Measures ANOVA within Asian Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 2.62$ | .107 | .01 |
| Emotional Expression | $F(2.78,1247.53) = 1039.94$ | < .001 | .70 |
| Image Type*Emotional Expression | $F(2.69,1207.41) = 69.82$ | < .001 | .14 |

*Note.* N = 450.

**Table S8**

*Accuracy: Results for the 2 × 4 Repeated Measures ANOVA within Black Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 1.66$ | .198 | .004 |
| Emotional Expression | $F(2.21,990.42) = 1653.32$ | < .001 | .79 |
| Image Type*Emotional Expression | $F(2.36,1060.60) = 531.31$ | < .001 | .54 |

*Note.* N = 450.

**Table S9**

*Accuracy: Results for the 2 × 4 Repeated Measures ANOVA within Black Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 19.26$ | < .001 | .04 |
| Emotional Expression | $F(2.26,1013.08) = 2906.36$ | < .001 | .87 |
| Image Type*Emotional Expression | $F(2.27,1017.61) = 150.16$ | < .001 | .25 |

*Note.* N = 450.

**Table S10**

*Accuracy: Results for the 2 × 4 Repeated Measures ANOVA within White Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 278.91$ | < .001 | .38 |
| Emotional Expression | $F(2.90,1301.77) = 1315.16$ | < .001 | .75 |
| Image Type*Emotional Expression | $F(2.85,1277.66) = 798.85$ | < .001 | .64 |

*Note.* N = 450.

**Table S11**

*Accuracy: Results for the 2 × 4 Repeated Measures ANOVA within White Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 17.25$ | < .001 | .04 |
| Emotional Expression | $F(2.66,1194.29) = 1703.13$ | < .001 | .79 |
| Image Type*Emotional Expression | $F(2.72,1218.96) = 86.48$ | < .001 | .16 |

*Note.* N = 450.

**Table S12**

*Accuracy: t-tests*

| Stimulus Race/Gender | Emotional Expression | Image Type | Image Type | *t*-value | *p*-value | Cohen's *d* |
|---|---|---|---|---|---|---|
| Asian Man | Anger | AI-generated | Real | -16.71 | <.001 | -0.79 |
| | Fear | AI-generated | Real | 1.75 | 0.085 | 0.08 |
| | Happiness | AI-generated | Real | -22.71 | <.001 | -1.07 |
| | Sadness | AI-generated | Real | 18.38 | <.001 | 0.87 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 0.47 | 1 | 0.02 |
| | | Anger | Happiness | -23.76 | <.001 | -1.12 |
| | | Anger | Sadness | -7.74 | <.001 | -0.36 |
| | | Fear | Happiness | -22.89 | <.001 | -1.08 |
| | | Fear | Sadness | -8.11 | <.001 | -0.38 |
| | | Happiness | Sadness | 12.90 | <.001 | 0.61 |
| | Real | Anger | Fear | 16.79 | <.001 | 0.79 |
| | | Anger | Happiness | -25.82 | <.001 | -1.22 |
| | | Anger | Sadness | 25.35 | <.001 | 1.20 |
| | | Fear | Happiness | -58.31 | <.001 | -2.75 |
| | | Fear | Sadness | 8.62 | <.001 | 0.41 |
| | | Happiness | Sadness | 87.00 | <.001 | 4.10 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| Asian Woman | Anger | AI-generated | Real | -11.06 | <.001 | -0.52 |
| | Fear | AI-generated | Real | 1.50 | 0.122 | 0.07 |
| | Happiness | AI-generated | Real | -1.73 | 0.067 | -0.08 |
| | Sadness | AI-generated | Real | 8.71 | <.001 | 0.41 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 7.59 | <.001 | 0.36 |
| | | Anger | Happiness | -31.50 | <.001 | -1.48 |
| | | Anger | Sadness | 1.80 | 0.42 | 0.08 |
| | | Fear | Happiness | -35.63 | <.001 | -1.68 |
| | | Fear | Sadness | -5.61 | <.001 | -0.26 |
| | | Happiness | Sadness | 33.50 | <.001 | 1.58 |
| | Real | Anger | Fear | 17.53 | <.001 | 0.83 |
| | | Anger | Happiness | -17.75 | <.001 | -0.84 |
| | | Anger | Sadness | 19.24 | <.001 | 0.91 |
| | | Fear | Happiness | -41.13 | <.001 | -1.94 |
| | | Fear | Sadness | -0.47 | 1 | -0.02 |
| | | Happiness | Sadness | 55.45 | <.001 | 2.61 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

*Accuracy: t-tests cont.*

| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| Black Man | Anger | AI-generated | Real | -8.211 | <.001 | -0.387 |
| | Fear | AI-generated | Real | -13.143 | <.001 | -0.620 |
| | Happiness | AI-generated | Real | -17.143 | <.001 | -0.808 |
| | Sadness | AI-generated | Real | 30.824 | <.001 | 1.453 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 26.706 | <.001 | 1.259 |
| | | Anger | Happiness | -14.650 | <.001 | -0.691 |
| | | Anger | Sadness | -5.538 | <.001 | -0.261 |
| | | Fear | Happiness | -53.357 | <.001 | -2.515 |
| | | Fear | Sadness | -37.438 | <.001 | -1.765 |
| | | Happiness | Sadness | 7.450 | <.001 | 0.351 |
| | Real | Anger | Fear | 32.375 | <.001 | 1.526 |
| | | Anger | Happiness | -25.200 | <.001 | -1.188 |
| | | Anger | Sadness | 35.733 | <.001 | 1.684 |
| | | Fear | Happiness | -112.000 | <.001 | -5.280 |
| | | Fear | Sadness | 1.800 | 0.403 | 0.085 |
| | | Happiness | Sadness | 130.429 | <.001 | 6.148 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| Black Woman | Anger | AI-generated | Real | 14.056 | <.001 | 0.663 |
| | Fear | AI-generated | Real | -11.167 | <.001 | -0.526 |
| | Happiness | AI-generated | Real | -7.889 | <.001 | -0.372 |
| | Sadness | AI-generated | Real | -0.077 | 0.911 | -0.004 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 45.733 | <.001 | 2.156 |
| | | Anger | Happiness | -12.529 | <.001 | -0.591 |
| | | Anger | Sadness | 24.647 | <.001 | 1.162 |
| | | Fear | Happiness | -89.900 | <.001 | -4.238 |
| | | Fear | Sadness | -26.700 | <.001 | -1.259 |
| | | Happiness | Sadness | 52.667 | <.001 | 2.483 |
| | Real | Anger | Fear | 24.467 | <.001 | 1.153 |
| | | Anger | Happiness | -38.286 | <.001 | -1.805 |
| | | Anger | Sadness | 10.313 | <.001 | 0.486 |
| | | Fear | Happiness | -129.000 | <.001 | -6.081 |
| | | Fear | Sadness | -16.750 | <.001 | -0.790 |
| | | Happiness | Sadness | 58.417 | <.001 | 2.754 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

*Accuracy: t-tests cont.*

| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| White Man | Anger | AI-generated | Real | -30.00 | <.001 | -1.41 |
| | Fear | AI-generated | Real | -18.71 | <.001 | -0.88 |
| | Happiness | AI-generated | Real | -18.86 | <.001 | -0.89 |
| | Sadness | AI-generated | Real | 30.29 | <.001 | 1.43 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 24.69 | <.001 | 1.16 |
| | | Anger | Happiness | -17.88 | <.001 | -0.84 |
| | | Anger | Sadness | -11.35 | <.001 | -0.54 |
| | | Fear | Happiness | -41.60 | <.001 | -1.96 |
| | | Fear | Sadness | -32.18 | <.001 | -1.52 |
| | | Happiness | Sadness | 4.05 | <.001 | 0.19 |
| | Real | Anger | Fear | 33.87 | <.001 | 1.60 |
| | | Anger | Happiness | -10.82 | <.001 | -0.51 |
| | | Anger | Sadness | 61.50 | <.001 | 2.90 |
| | | Fear | Happiness | -44.79 | <.001 | -2.11 |
| | | Fear | Sadness | 15.33 | <.001 | 0.72 |
| | | Happiness | Sadness | 95.11 | <.001 | 4.48 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| White Woman | Anger | AI-generated | Real | -7.87 | <.001 | -0.37 |
| | Fear | AI-generated | Real | 7.12 | <.001 | 0.34 |
| | Happiness | AI-generated | Real | -5.50 | <.001 | -0.26 |
| | Sadness | AI-generated | Real | 9.94 | <.001 | 0.47 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | -1.94 | 0.311 | -0.09 |
| | | Anger | Happiness | -51.08 | <.001 | -2.41 |
| | | Anger | Sadness | -11.44 | <.001 | -0.54 |
| | | Fear | Happiness | -39.38 | <.001 | -1.86 |
| | | Fear | Sadness | -8.22 | <.001 | -0.39 |
| | | Happiness | Sadness | 37.00 | <.001 | 1.74 |
| | Real | Anger | Fear | 12.75 | <.001 | 0.60 |
| | | Anger | Happiness | -49.25 | <.001 | -2.32 |
| | | Anger | Sadness | 5.53 | <.001 | 0.26 |
| | | Fear | Happiness | -66.33 | <.001 | -3.13 |
| | | Fear | Sadness | -7.33 | <.001 | -0.35 |
| | | Happiness | Sadness | 57.08 | <.001 | 2.69 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

**Table S13**

*Intensity: Results for the 2 × 2 × 3 × 4 Repeated Measures ANOVA*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 1340.68$ | < .001 | .75 |
| Simulus Gender | $F(1,449) = 262.95$ | < .001 | .37 |
| Stimulus Race | $F(2,898) = 309.84$ | < .001 | .41 |
| Emotional Expression | $F(2.92,1311.70) = 854.11$ | < .001 | .66 |
| Image Type*Stimulus Gender | $F(1,449) = 554.27$ | < .001 | .55 |
| Image Type*Stimulus Race | $F(2,898) = 37.83$ | < .001 | .08 |
| Image Type*Emotional Expression | $F(2.74,1230.84) = 774.18$ | < .001 | .63 |
| Stimulus Gender*Stimulus Race | $F(2,898) = 613.04$ | < .001 | .58 |
| Stimulus Gender*Emotional Expression | $F(2.90,1299.68) = 187.81$ | < .001 | .30 |
| Stimulus Race*Emotional Expression | $F(5.24,2353.52) = 337.37$ | < .001 | .43 |
| Image Type *Stimulus Gender*Stimulus Race | $F(1.98,890.25) = 271.12$ | < .001 | .38 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.74,1229.25) = 901.27$ | < .001 | .67 |
| Image Type*Stimulus Race*Emotional Expression | $F(5.55,2492.16) = 137.99$ | < .001 | .24 |
| Stimulus Gender*Stimulus Race*Emotional Expression | $F(5.64,2533.17) = 130.14$ | < .001 | .23 |
| Image Type*Stimulus Gender*Stimulus Race*Emotional Expression | $F(5.53,2483.39) = 144.69$ | < .001 | .24 |

**Table S14**

*Intensity: Results for the 2 × 2 × 4 Repeated Measures ANOVA within Asian Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 878.38$ | < .001 | .66 |
| Simulus Gender | $F(1,449) = 1253.07$ | < .001 | .74 |
| Emotional Expression | $F(2.88,1291.78) = 375.42$ | < .001 | .46 |
| Image Type*Stimulus Gender | $F(1, 449) = 1.00$ | .317 | .002 |
| Image Type*Emotional Expression | $F(2.77,1244.17) = 344.75$ | < .001 | .43 |
| Stimulus Gender*Emotional Expression | $F(2.86,1285.35) = 2.88$ | .037 | .006 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.87,1286.38) = 179.68$ | < .001 | .29 |

**Table S15**

*Intensity: Results for the 2 × 2 × 4 Repeated Measures ANOVA within Black Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 889.89$ | < .001 | .67 |
| Simulus Gender | $F(1,449) = 19.42$ | < .001 | .04 |
| Emotional Expression | $F(3,1347) = 1211.85$ | < .001 | .73 |
| Image Type*Stimulus Gender | $F(1, 449) = 87.33$ | < .001 | .16 |
| Image Type*Emotional Expression | $F(2.83,1271.20) = 642.33$ | < .001 | .59 |
| Stimulus Gender*Emotional Expression | $F(2.79,1250.53) = 72.40$ | < .001 | .14 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.74,1230.44) = 523.84$ | < .001 | .54 |

**Table S16**

*Intensity: Results for the 2 × 2 × 4 Repeated Measures ANOVA within White Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 1223.33$ | < .001 | .73 |
| Simulus Gender | $F(1,449) = 3.21$ | .07 | .007 |
| Emotional Expression | $F(2.92,1310.48) = 562.03$ | < .001 | .56 |
| Image Type*Stimulus Gender | $F(1, 449) = 1115.04$ | < .001 | .71 |
| Image Type*Emotional Expression | $F(2.82,1264.19) = 633.34$ | < .001 | .59 |
| Stimulus Gender*Emotional Expression | $F(2.89,1297.13) = 400.33$ | < .001 | .47 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.76,1237.28) = 646.93$ | < .001 | .59 |

**Table S17**

*Intensity: Results for the 2 × 4 Repeated Measures ANOVA within Asian Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 716.35$ | < .001 | .62 |
| Emotional Expression | $F(2.92,1311.20) = 275.34$ | < .001 | .38 |
| Image Type*Emotional Expression | $F(2.88,1293.70) = 461.72$ | < .001 | .51 |

**Table S18**

*Intensity: Results for the 2 × 4 Repeated Measures ANOVA within Asian Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 580.83$ | < .001 | .56 |
| Emotional Expression | $F(2.85,1280.73) = 294.63$ | < .001 | .40 |
| Image Type*Emotional Expression | $F(2.68,1202.74) = 59.69$ | < .001 | .12 |

**Table S19**

*Intensity: Results for the 2 × 4 Repeated Measures ANOVA within Black Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 856.51$ | < .001 | .66 |
| Emotional Expression | $F(2.88,1291.64) = 613.47$ | < .001 | .58 |
| Image Type*Emotional Expression | $F(2.83,1270.24) = 779.44$ | < .001 | .63 |

**Table S20**

*Intensity: Results for the 2 × 4 Repeated Measures ANOVA within Black Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 466.58$ | < .001 | .51 |
| Emotional Expression | $F(3,1347) = 1305.40$ | < .001 | .74 |
| Image Type*Emotional Expression | $F(2.78,1248.15) = 358.82$ | < .001 | .44 |

**Table S21**

*Intensity: Results for the 2 × 4 Repeated Measures ANOVA within White Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 2146.19$ | < .001 | .83 |
| Emotional Expression | $F(2.92,1311.59) = 429.60$ | < .001 | .49 |
| Image Type*Emotional Expression | $F(2.68,1203.76) = 999.51$ | < .001 | .69 |

**Table S22**

*Intensity: Results for the 2 × 4 Repeated Measures ANOVA within White Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 219.36$ | < .001 | .33 |
| Emotional Expression | $F(2.85,1278.24) = 602.78$ | < .001 | .57 |
| Image Type*Emotional Expression | $F(2.83,1270.68) = 217.45$ | < .001 | .33 |

**Table S23**

*Intensity: t-tests*

| Stimulus Race/Gender | Emotional Expression | Image Type | Image Type | *t*-value | *p*-value | Cohen's *d* |
|---|---|---|---|---|---|---|
| Asian Man | Anger | AI-generated | Real | -18.28 | <.001 | -0.86 |
| | Fear | AI-generated | Real | -14.44 | <.001 | -0.68 |
| | Happiness | AI-generated | Real | -45.11 | <.001 | -2.13 |
| | Sadness | AI-generated | Real | 16.62 | <.001 | 0.78 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 1.92 | 0.339 | 0.09 |
| | | Anger | Happiness | 1.70 | 0.523 | 0.08 |
| | | Anger | Sadness | -1.26 | 1 | -0.06 |
| | | Fear | Happiness | -0.29 | 1 | -0.01 |
| | | Fear | Sadness | -3.39 | 0.005 | -0.16 |
| | | Happiness | Sadness | -2.98 | 0.018 | -0.14 |
| | Real | Anger | Fear | 3.59 | 0.002 | 0.17 |
| | | Anger | Happiness | -13.35 | <.001 | -0.63 |
| | | Anger | Sadness | 31.77 | <.001 | 1.50 |
| | | Fear | Happiness | -15.67 | <.001 | -0.74 |
| | | Fear | Sadness | 26.03 | <.001 | 1.23 |
| | | Happiness | Sadness | 52.60 | <.001 | 2.48 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| Asian Woman | Anger | AI-generated | Real | -18.57 | <.001 | -0.88 |
| | Fear | AI-generated | Real | -12.49 | <.001 | -0.59 |
| | Happiness | AI-generated | Real | -26.22 | <.001 | -1.24 |
| | Sadness | AI-generated | Real | -3.87 | <.001 | -0.18 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 3.20 | 0.009 | 0.15 |
| | | Anger | Happiness | -7.30 | <.001 | -0.34 |
| | | Anger | Sadness | 10.41 | <.001 | 0.49 |
| | | Fear | Happiness | -9.54 | <.001 | -0.45 |
| | | Fear | Sadness | 6.00 | <.001 | 0.28 |
| | | Happiness | Sadness | 18.00 | <.001 | 0.85 |
| | Real | Anger | Fear | 5.35 | <.001 | 0.25 |
| | | Anger | Happiness | -6.40 | <.001 | -0.30 |
| | | Anger | Sadness | 22.56 | <.001 | 1.06 |
| | | Fear | Happiness | -11.16 | <.001 | -0.53 |
| | | Fear | Sadness | 15.13 | <.001 | 0.71 |
| | | Happiness | Sadness | 32.44 | <.001 | 1.53 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

*Intensity: t-tests cont.*

| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| Black Man | Anger | AI-generated | Real | -15.63 | <.001 | -0.74 |
| | Fear | AI-generated | Real | -26.23 | <.001 | -1.24 |
| | Happiness | AI-generated | Real | -50.92 | <.001 | -2.40 |
| | Sadness | AI-generated | Real | 19.69 | <.001 | 0.93 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 28.38 | <.001 | 1.34 |
| | | Anger | Happiness | 6.49 | <.001 | 0.31 |
| | | Anger | Sadness | -0.42 | 1 | -0.02 |
| | | Fear | Happiness | -24.27 | <.001 | -1.14 |
| | | Fear | Sadness | -27.97 | <.001 | -1.32 |
| | | Happiness | Sadness | -6.57 | <.001 | -0.31 |
| | Real | Anger | Fear | 11.91 | <.001 | 0.56 |
| | | Anger | Happiness | -22.11 | <.001 | -1.04 |
| | | Anger | Sadness | 36.98 | <.001 | 1.74 |
| | | Fear | Happiness | -29.54 | <.001 | -1.39 |
| | | Fear | Sadness | 18.34 | <.001 | 0.86 |
| | | Happiness | Sadness | 67.48 | <.001 | 3.18 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| Black Woman | Anger | AI-generated | Real | 9.35 | <.001 | 0.44 |
| | Fear | AI-generated | Real | -18.20 | <.001 | -0.86 |
| | Happiness | AI-generated | Real | -37.11 | <.001 | -1.75 |
| | Sadness | AI-generated | Real | -11.76 | <.001 | -0.55 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 44.34 | <.001 | 2.09 |
| | | Anger | Happiness | 7.57 | <.001 | 0.36 |
| | | Anger | Sadness | 33.81 | <.001 | 1.59 |
| | | Fear | Happiness | -40.37 | <.001 | -1.90 |
| | | Fear | Sadness | -18.62 | <.001 | -0.88 |
| | | Happiness | Sadness | 25.65 | <.001 | 1.21 |
| | Real | Anger | Fear | 11.78 | <.001 | 0.56 |
| | | Anger | Happiness | -33.24 | <.001 | -1.57 |
| | | Anger | Sadness | 9.39 | <.001 | 0.44 |
| | | Fear | Happiness | -45.50 | <.001 | -2.14 |
| | | Fear | Sadness | -3.56 | 0.003 | -0.17 |
| | | Happiness | Sadness | 46.31 | <.001 | 2.18 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

*Intensity: t-tests cont.*

| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| White Man | Anger | AI-generated | Real | -44.23 | <.001 | -2.09 |
| | Fear | AI-generated | Real | -37.76 | <.001 | -1.78 |
| | Happiness | AI-generated | Real | -50.13 | <.001 | -2.36 |
| | Sadness | AI-generated | Real | 17.08 | <.001 | 0.80 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 16.69 | <.001 | 0.79 |
| | | Anger | Happiness | 1.35 | 1 | 0.06 |
| | | Anger | Sadness | -8.03 | <.001 | -0.38 |
| | | Fear | Happiness | -14.32 | <.001 | -0.68 |
| | | Fear | Sadness | -22.84 | <.001 | -1.08 |
| | | Happiness | Sadness | -8.92 | <.001 | -0.42 |
| | Real | Anger | Fear | 14.66 | <.001 | 0.69 |
| | | Anger | Happiness | 3.13 | 0.01 | 0.15 |
| | | Anger | Sadness | 53.55 | <.001 | 2.52 |
| | | Fear | Happiness | -12.02 | <.001 | -0.57 |
| | | Fear | Sadness | 32.48 | <.001 | 1.53 |
| | | Happiness | Sadness | 54.85 | <.001 | 2.59 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| White Woman | Anger | AI-generated | Real | -1.93 | 0.052 | -0.09 |
| | Fear | AI-generated | Real | -0.61 | 0.545 | -0.03 |
| | Happiness | AI-generated | Real | -36.82 | <.001 | -1.74 |
| | Sadness | AI-generated | Real | -2.95 | 0.003 | -0.14 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | -0.03 | 1 | 0.00 |
| | | Anger | Happiness | -11.50 | <.001 | -0.54 |
| | | Anger | Sadness | 5.18 | <.001 | 0.24 |
| | | Fear | Happiness | -9.60 | <.001 | -0.45 |
| | | Fear | Sadness | 5.03 | <.001 | 0.24 |
| | | Happiness | Sadness | 17.41 | <.001 | 0.82 |
| | Real | Anger | Fear | 1.23 | 1 | 0.06 |
| | | Anger | Happiness | -46.49 | <.001 | -2.19 |
| | | Anger | Sadness | 4.17 | <.001 | 0.20 |
| | | Fear | Happiness | -45.20 | <.001 | -2.13 |
| | | Fear | Sadness | 2.85 | 0.029 | 0.13 |
| | | Happiness | Sadness | 47.67 | <.001 | 2.25 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

**Table S24**

*Saliency: Results for the 2 × 2 × 3 × 4 Repeated Measures ANOVA*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 86.05$ | < .001 | .16 |
| Simulus Gender | $F(1,449) = 103.95$ | < .001 | .19 |
| Stimulus Race | $F(2,898) = 76.29$ | < .001 | .15 |
| Emotional Expression | $F(2.80,1258.79) = 3082.81$ | < .001 | .87 |
| Image Type*Stimulus Gender | $F(1,449) = 279.54$ | < .001 | .38 |
| Image Type*Stimulus Race | $F(2,898) = 46.22$ | < .001 | .09 |
| Image Type*Emotional Expression | $F(2.85,1281.20) = 570.57$ | < .001 | .56 |
| Stimulus Gender*Stimulus Race | $F(2,898) = 465.78$ | < .001 | .51 |
| Stimulus Gender*Emotional Expression | $F(2.88,1293.38) = 123.63$ | < .001 | .22 |
| Stimulus Race*Emotional Expression | $F(5.56,2496.17) = 274.46$ | < .001 | .38 |
| Image Type *Stimulus Gender*Stimulus Race | $F(2,898) = 96.14$ | < .001 | .18 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.69,1206.50) = 700.58$ | < .001 | .61 |
| Image Type*Stimulus Race*Emotional Expression | $F(5.72,2569.73) = 182.63$ | < .001 | .29 |
| Stimulus Gender*Stimulus Race*Emotional Expression | $F(5.71,2565.27) = 150.89$ | < .001 | .25 |
| Image Type*Stimulus Gender*Stimulus Race*Emotional Expression | $F(5.71,2564.41) = 112.91$ | < .001 | .20 |

**Table S25**

*Saliency: Results for the 2 × 2 × 4 Repeated Measures ANOVA within Asian Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 124.20$ | < .001 | .22 |
| Simulus Gender | $F(1,449) = 824.33$ | < .001 | .65 |
| Emotional Expression | $F(2.89,1296.20) = 1732.00$ | < .001 | .79 |
| Image Type*Stimulus Gender | $F(1, 449) = 44.48$ | < .001 | .09 |
| Image Type*Emotional Expression | $F(2.95,1322.01) = 292.56$ | < .001 | .40 |
| Stimulus Gender*Emotional Expression | $F(3,1347) = 13.88$ | < .001 | .03 |
| Image Type*Stimulus Gender*Emotional Expression | $F(3,1347) = 164.71$ | < .001 | .27 |

**Table S26**

*Saliency: Results for the 2 × 2 × 4 Repeated Measures ANOVA within Black Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 1.30$ | .256 | .003 |
| Simulus Gender | $F(1,449) = 11.84$ | < .001 | .03 |
| Emotional Expression | $F(2.67,1199.49) = 2744.97$ | < .001 | .86 |
| Image Type*Stimulus Gender | $F(1, 449) = 15.61$ | < .001 | .03 |
| Image Type*Emotional Expression | $F(2.85,1281.48) = 374.67$ | < .001 | .46 |
| Stimulus Gender*Emotional Expression | $F(2.74,1229.71) = 60.43$ | < .001 | .12 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.61,1171.55) = 524.38$ | < .001 | .54 |

**Table S27**

*Saliency: Results for the 2 × 2 × 4 Repeated Measures ANOVA within White Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 84.16$ | < .001 | .16 |
| Simulus Gender | $F(1,449) = 74.52$ | < .001 | .14 |
| Emotional Expression | $F(2.92,1310.89) = 2172.15$ | < .001 | .83 |
| Image Type*Stimulus Gender | $F(1, 449) = 496.85$ | < .001 | .53 |
| Image Type*Emotional Expression | $F(2.85,1280.53) = 627.54$ | < .001 | .58 |
| Stimulus Gender*Emotional Expression | $F(3,1347) = 374.90$ | < .001 | .46 |
| Image Type*Stimulus Gender*Emotional Expression | $F(2.87,1288.89) = 382.15$ | < .001 | .46 |

**Table S28**

*Saliency: Results for the 2 × 4 Repeated Measures ANOVA within Asian Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 143.39$ | < .001 | .24 |
| Emotional Expression | $F(2.90,1299.75) = 1029.98$ | < .001 | .70 |
| Image Type*Emotional Expression | $F(2.94,1322.68) = 352.91$ | < .001 | .44 |

**Table S29**

*Saliency: Results for the 2 × 4 Repeated Measures ANOVA within Asian Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 28.38$ | < .001 | .06 |
| Emotional Expression | $F(2.90,1302.61) = 1532.34$ | < .001 | .77 |
| Image Type*Emotional Expression | $F(2.88,1292.50) = 94.94$ | < .001 | .18 |

**Table S30**

*Saliency: Results for the 2 × 4 Repeated Measures ANOVA within Black Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 11.00$ | < .001 | .02 |
| Emotional Expression | $F(2.68,1202.74) = 1437.36$ | < .001 | .76 |
| Image Type*Emotional Expression | $F(2.73,1225.81) = 582.48$ | < .001 | .57 |

**Table S31**

*Saliency: Results for the 2 × 4 Repeated Measures ANOVA within Black Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 2.66$ | .104 | .006 |
| Emotional Expression | $F(2.73,1225.72) = 2736.59$ | < .001 | .86 |
| Image Type*Emotional Expression | $F(2.73,1225.91) = 199.02$ | < .001 | .31 |

**Table S32**

*Saliency: Results for the 2 × 4 Repeated Measures ANOVA within White Men's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 369.31$ | < .001 | .45 |
| Emotional Expression | $F(2.94,1321.49) = 1176.75$ | < .001 | .72 |
| Image Type*Emotional Expression | $F(2.80,1256.63) = 793.46$ | < .001 | .64 |

**Table S33**

*Saliency: Results for the 2 × 4 Repeated Measures ANOVA within White Women's Faces*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | $F(1,449) = 45.43$ | < .001 | .09 |
| Emotional Expression | $F(2.86,1284.24) = 2236.67$ | < .001 | .83 |
| Image Type*Emotional Expression | $F(2.85,1280.89) = 104.82$ | < .001 | .19 |

**Table S34**

*Saliency: t-tests*

| Stimulus Race/Gender | Emotional Expression | Image Type | Image Type | *t*-value | *p*-value | Cohen's *d* |
|---|---|---|---|---|---|---|
| Asian Man | Anger | AI-generated | Real | -18.23 | <.001 | -0.86 |
| | Fear | AI-generated | Real | -3.67 | <.001 | -0.17 |
| | Happiness | AI-generated | Real | -21.62 | <.001 | -1.02 |
| | Sadness | AI-generated | Real | 19.07 | <.001 | 0.90 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | -1.00 | 1 | -0.05 |
| | | Anger | Happiness | -23.00 | <.001 | -1.08 |
| | | Anger | Sadness | -8.50 | <.001 | -0.40 |
| | | Fear | Happiness | -22.13 | <.001 | -1.04 |
| | | Fear | Sadness | -8.20 | <.001 | -0.39 |
| | | Happiness | Sadness | 13.59 | <.001 | 0.64 |
| | Real | Anger | Fear | 12.86 | <.001 | 0.61 |
| | | Anger | Happiness | -34.25 | <.001 | -1.61 |
| | | Anger | Sadness | 26.29 | <.001 | 1.24 |
| | | Fear | Happiness | -59.20 | <.001 | -2.79 |
| | | Fear | Sadness | 17.09 | <.001 | 0.81 |
| | | Happiness | Sadness | 77.90 | <.001 | 3.67 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| Asian Woman | Anger | AI-generated | Real | -15.18 | <.001 | -0.72 |
| | Fear | AI-generated | Real | -2.33 | 0.022 | -0.11 |
| | Happiness | AI-generated | Real | -1.00 | 0.305 | -0.05 |
| | Sadness | AI-generated | Real | 8.45 | <.001 | 0.40 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 6.58 | <.001 | 0.31 |
| | | Anger | Happiness | -40.27 | <.001 | -1.90 |
| | | Anger | Sadness | 1.09 | 1 | 0.05 |
| | | Fear | Happiness | -43.50 | <.001 | -2.05 |
| | | Fear | Sadness | -5.15 | <.001 | -0.24 |
| | | Happiness | Sadness | 41.36 | <.001 | 1.95 |
| | Real | Anger | Fear | 16.77 | <.001 | 0.79 |
| | | Anger | Happiness | -23.83 | <.001 | -1.12 |
| | | Anger | Sadness | 22.58 | <.001 | 1.06 |
| | | Fear | Happiness | -45.82 | <.001 | -2.16 |
| | | Fear | Sadness | 4.82 | <.001 | 0.23 |
| | | Happiness | Sadness | 55.80 | <.001 | 2.63 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

*Saliency: t-tests cont.*

| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| Black Man | Anger | AI-generated | Real | -7.93 | <.001 | -0.37 |
| | Fear | AI-generated | Real | -20.80 | <.001 | -0.98 |
| | Happiness | AI-generated | Real | -16.85 | <.001 | -0.79 |
| | Sadness | AI-generated | Real | 30.60 | <.001 | 1.44 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 26.27 | <.001 | 1.24 |
| | | Anger | Happiness | -15.59 | <.001 | -0.73 |
| | | Anger | Sadness | -5.60 | <.001 | -0.26 |
| | | Fear | Happiness | -43.93 | <.001 | -2.07 |
| | | Fear | Sadness | -33.80 | <.001 | -1.59 |
| | | Happiness | Sadness | 8.44 | <.001 | 0.40 |
| | Real | Anger | Fear | 21.21 | <.001 | 1.00 |
| | | Anger | Happiness | -37.30 | <.001 | -1.76 |
| | | Anger | Sadness | 35.23 | <.001 | 1.66 |
| | | Fear | Happiness | -66.90 | <.001 | -3.15 |
| | | Fear | Sadness | 14.64 | <.001 | 0.69 |
| | | Happiness | Sadness | 103.75 | <.001 | 4.89 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| Black Woman | Anger | AI-generated | Real | 17.31 | <.001 | 0.82 |
| | Fear | AI-generated | Real | -12.89 | <.001 | -0.61 |
| | Happiness | AI-generated | Real | -7.63 | <.001 | -0.36 |
| | Sadness | AI-generated | Real | -0.92 | 0.346 | -0.04 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 44.08 | <.001 | 2.08 |
| | | Anger | Happiness | -18.15 | <.001 | -0.86 |
| | | Anger | Sadness | 25.21 | <.001 | 1.19 |
| | | Fear | Happiness | -73.55 | <.001 | -3.47 |
| | | Fear | Sadness | -22.00 | <.001 | -1.04 |
| | | Happiness | Sadness | 49.08 | <.001 | 2.31 |
| | Real | Anger | Fear | 17.85 | <.001 | 0.84 |
| | | Anger | Happiness | -43.50 | <.001 | -2.05 |
| | | Anger | Sadness | 9.00 | <.001 | 0.42 |
| | | Fear | Happiness | -83.78 | <.001 | -3.95 |
| | | Fear | Sadness | -10.45 | <.001 | -0.49 |
| | | Happiness | Sadness | 58.09 | <.001 | 2.74 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

*Saliency: t-tests cont.*

| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
|---|---|---|---|---|---|---|
| White Man | Anger | AI-generated | Real | -28.42 | <.001 | -1.34 |
| | Fear | AI-generated | Real | -25.00 | <.001 | -1.18 |
| | Happiness | AI-generated | Real | -19.31 | <.001 | -0.91 |
| | Sadness | AI-generated | Real | 31.21 | <.001 | 1.47 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | 20.08 | <.001 | 0.95 |
| | | Anger | Happiness | -19.80 | <.001 | -0.93 |
| | | Anger | Sadness | -12.50 | <.001 | -0.59 |
| | | Fear | Happiness | -35.80 | <.001 | -1.69 |
| | | Fear | Sadness | -29.40 | <.001 | -1.39 |
| | | Happiness | Sadness | 5.71 | <.001 | 0.27 |
| | Real | Anger | Fear | 23.42 | <.001 | 1.10 |
| | | Anger | Happiness | -20.70 | <.001 | -0.98 |
| | | Anger | Sadness | 44.38 | <.001 | 2.09 |
| | | Fear | Happiness | -48.80 | <.001 | -2.30 |
| | | Fear | Sadness | 24.67 | <.001 | 1.16 |
| | | Happiness | Sadness | 87.11 | <.001 | 4.11 |
| **Stimulus Race/Gender** | **Emotional Expression** | **Image Type** | **Image Type** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| White Woman | Anger | AI-generated | Real | -5.18 | <.001 | -0.24 |
| | Fear | AI-generated | Real | 6.30 | <.001 | 0.30 |
| | Happiness | AI-generated | Real | -5.71 | <.001 | -0.27 |
| | Sadness | AI-generated | Real | 14.00 | <.001 | 0.66 |
| | **Image Type** | **Emotional Expression** | **Emotional Expression** | ***t*-value** | ***p*-value** | **Cohen's *d*** |
| | AI-generated | Anger | Fear | -1.15 | 1 | -0.05 |
| | | Anger | Happiness | -52.00 | <.001 | -2.45 |
| | | Anger | Sadness | -12.25 | <.001 | -0.58 |
| | | Fear | Happiness | -46.42 | <.001 | -2.19 |
| | | Fear | Sadness | -9.50 | <.001 | -0.45 |
| | | Happiness | Sadness | 35.42 | <.001 | 1.67 |
| | Real | Anger | Fear | 9.64 | <.001 | 0.45 |
| | | Anger | Happiness | -61.56 | <.001 | -2.90 |
| | | Anger | Sadness | 6.58 | <.001 | 0.31 |
| | | Fear | Happiness | -82.50 | <.001 | -3.89 |
| | | Fear | Sadness | -2.70 | 0.051 | -0.13 |
| | | Happiness | Sadness | 70.33 | <.001 | 3.32 |

*Note.* N = 450. All pairwise comparisons were Bonferroni adjusted in SPSS.

Study 3 has been submitted for publication.

## 4. STUDY 3: Social Attributions Given to Happy and Angry AI-Generated and Real Faces: Is AI Reproducing Race and Gender Biases?

Megan Lawrence*, Emily Carriere, & Charles A. Collin

### 4.1. Abstract

We assign social attributions to others based on inferences of their emotional states. This guides our behaviours towards them. With artificial intelligence (AI) generated images becoming ubiquitous, understanding how they represent the diversity of real faces is important. Previous research finds that AI-generators replicate and amplify existing stereotypes. Our study used mixed measures to examine whether AI-generated faces evoke similar responses to real photographs and how knowledge about the images' origins influence these judgements. N=81 undergraduates viewed AI-generated images and real photographs portraying different genders (men, women), races (Asian, Black, White), and expressions (angry, happy). They were given either correct, incorrect, or no information regarding which faces were real vs. AI, and they then provided their first impressions. Responses were coded on four dependent variables (valence, arousal, warmth, dominance) and examined quantitatively. Results indicate that the AI-generator, an implementation of Stable Diffusion, reflected/amplified stereotypes, particularly about Black and Asian women.

## 4.2. Introduction

We assign social attributes to others when we meet them based on quick judgements. These judgements may be informed by facial expressions, perceived attractiveness, race, gender, age, and other similar factors. The first impressions we make of individuals can have a lasting impact because social attributes are generally considered to be permanent and inherent. In subsequent interactions, our judgements of a person's social attributes can be confirmed or challenged, or may persist despite information contrary to our initial judgement (Ames & Bianchi, 2008; Todorov et al., 2015; Zebrowitz, 2017). For example, Ames and Bianchi (2008) found that agreeableness judgements tend to persist beyond first impressions.

Social attributions can be affected by a number of different observer and model characteristics (Campbell et al., 2010; Cortes et al., 2019; Kiiski et al., 2016; Olivola et al., 2014; Zebrowitz, 2017; Zhang et al., 2020). For example, female faces have been rated as more approachable than male faces; and Caucasian faces are rated as more approachable than non-Caucasian faces by Caucasian participants (Campbell et al., 2010). In addition, participant age impacts ratings of attractiveness, likeability, competence, and trustworthiness (Cortes et al., 2019; Kiiski et al., 2016).

Research with AI-generated images has not studied social attributions extensively. Indeed, we are aware of only two studies that have examined this. One, by Nightingale and Farid (2022), found that trustworthiness ratings differ for real and AI-generated faces. More specifically, AI-generated faces that are assumed to be real were rated as more trustworthy than actual real faces. The authors hypothesize that this is because the AI-generated faces are more average, and average faces tend to be perceived as more trustworthy (Sofer et al., 2015). More recently, Liefooghe et al. (2023) examined trustworthiness of real faces when they are labelled as

real or AI-generated and found that those labelled as AI-generated were seen as less trustworthy, but only in the presence of faces labelled as real.

The current study aimed to increase our understanding of how social attributions are made to AI-generated images. This was done by allowing participants to provide open-ended assessments of the face images and then assessing, via quantitative analysis techniques, whether there were systematic differences between the AI-generated faces (AIFs) and the real face photographs (RFs), and how this interacted with participants' knowledge of whether the images were AI-generated vs. real.

How race and gender of the faces would interact with and impact judgements of the AI-generated faces was unclear. Previous research (AlDahoul et al., 2025; Assis & Moura; Bhardwaj et al., 2025; Bianchi et al., 2023; Ghosh & Caliskan, 2023; Jääskeläinen et al., 2025; Lawrence, Cimermanis, & Collin, 2025; Lawrence & Collin, 2025; Locke & Hogdon, 2024; Nightingale & Farid, 2022; Wu et al., 2025) has found that AI-generators are generally biased towards creating more realistic images of White men’s faces. Moreover, AI-generators tend to replicate, if not amplify, existing societal biases and stereotypes (AlDahoul et al., 2025; Assis & Moura, 2025; Bianchi et al., 2023 Chauhan et al., 2024; Gawali et al., 2024; Ghosh & Caliskan, 2023; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Mubashir, 2024; Park, 2024). For example, when using generic prompts for individuals in certain professions, the generator replicates gender and racial stereotypes. Because of this, when the same prompt is used to create White faces and faces of people of colour using AI image generators, we anticipate that we will see differences in the social judgements made about these faces. This is likely because AI-generators have more images to draw on for White faces in their training sets (Kärkkäinen & Joo, 2021), allowing them to create more average and attractive White faces, thus leading to

more positive social judgments. This effect will likely result in the differences between real and AI-generated faces being greater for images of people of colour than for those of White people.

The literature has also examined how processing of *presumed real* vs. *presumed fake* images differs. Eiserbeck et al. (2023), examined this and found that presumed fake positive faces had reduced intensity, slower evaluations, and more effortful evaluations. Importantly, all the faces in their study were actually real face photographs. Our study aims to expand upon these findings by examining how knowledge of whether an image is AI-generated or not (and whether this information given is accurate) will influence the social assessment of faces. Based on this and previous research examining trustworthiness judgements of RFs vs. AIFs (Liefooghe et al., 2023; Nightingale & Farid, 2022), we formulated the following pre-registered hypothesis (full pre-registration can be found here: https://osf.io/6yc4f/overview?view_only=2330c3dfd9f147c5994348cc09692628.

**H1:** AI generators disproportionately create average and attractive faces unless otherwise specified. We therefore predict that knowledge about whether a face is AI-generated or real will impact the social attributions assigned to that face.

Specifically, in the no-knowledge condition, AIFs will be assigned more positive attributes than RFs. This is based on previous research that found that AI-generated faces were overall rated as more trustworthy, which the authors posit could be due to their averageness (Nightingale & Farid, 2022). In the correct-knowledge condition, AIFs will be assigned less positive attributes than the RFs. Previous research found that faces labelled as AI-generated (even though they were real) were less trustworthy than correctly labelled real faces (Liefooghe et al., 2023), which may indicate a negative association with AI-generated faces in general. Finally, in the incorrect-knowledge condition, AIFs will be assigned the most positive social

attributes. Similarly to the rationales above, this hypothesis is based on work that found that AI-generated faces are rated as more trustworthy (Liefooghe et al., 2023; Nightingale & Farid, 2022), and that AI-generated faces perceived as real elicit higher social conformity (Tucciarelli et al., 2022).

We also used the following questions to guide exploration of topics which have yet to be examined in research with AI-generated faces.

**Q1:** Will race of the faces impact the social attributions assigned?

**Q2:** Will gender of the faces impact the social attributions assigned?

**Q3:** Will facial expression of the faces impact the social attributions assigned?

**Q4:** Will the factors of race, gender, and expression interact in determining the social attribution assigned to a face?

The lack of research in this area makes it difficult to form firm hypotheses about interactions of the above effects. However, we expect that they may interact with each other. For example, an AIF of an angry woman of colour will have more negative social attributes assigned to it than an AIF of a happy White man. However, it is not clear whether the different factors (race, gender, expression) will interact rather than producing additive effects.

## 4.3. Methods

### *4.3.1. Participants*

N = 88 undergraduate students from the University of Ottawa's Integrated System for Participation in Research were recruited. The total sample size was estimated using GPower (Faul et al., 2007) for F-test mixed design, independent groups effect, a power of .95 and a medium effect size of $f = .25$ (n = 189). Using equation 4c in Goulet & Cousineau (2019), an

adjusted sample size for repeated measurements was calculated. For the calculation, estimated n = 189, number of repeated measurements = 3, and estimated correlation = .14. From this calculation, it was estimated that we would need to recruit 82 participants. We deviated slightly from the pre-registration by recruiting an approximate additional 8% (rather than 20%). As per the pre-registration, data was first cleaned by removing those with incorrect catch trials and outlying durations based on Van Selst and Jolicoeur guidelines (1994; n = 7). After catch trial rejection and duration outlier rejection, our final sample size was N = 81 participants (see Supplemental Materials Table S1 for demographic breakdown of the total sample). Participants were compensated with course credit. This study was approved by the University of Ottawa Social Sciences and Humanities Research Ethics Board.

#### *4.3.2. Materials*

This study used AI-generated faces from our previous work (Lawrence & Collin, 2025), and real faces from the RADIATE Database (Conley et al., 2018; Tottenham et al., 2009). For details on the generation of the AI faces, please see our previous work (Lawrence & Collin, 2025). The faces could differ by gender (men or women), race (Asian, Black, or White), and emotional expression (anger or happiness). Three faces having each possible combination of these variations were presented.

#### *4.3.3. Procedure*

This study was conducted online using Qualtrics$^{TM}$ and involved participants making judgments of AIFs and RFs. They completed the demographics questionnaire, which asks about age, year of study, sex and gender, ethnic background, time spent in Canada, and visual acuity. Next, they completed the Prosopagnosia Index (Shah et al., 2015).

Participants were randomly assigned to one of three groups: 1) Correct Information, where each image was labelled, correctly identifying the image as either real or AI-generated, 2) Incorrect Information, where each image was labelled, incorrectly identifying the image as either real or AI-generated, and 3) No Information, where no images were labelled, giving no indication that there was a mix of real and AI-generated images. Some deception was involved in that participants were not told about this manipulation until after they had completed the task. All groups followed the same procedure, with the only difference being the prompts they were given as to whether the images they were viewing were real or AI-generated.

Participants were presented with 72 stimuli in random order, one at a time, and instructed to provide three to ten things they would think about them upon seeing them. Examples given in the instructions were personality, attractiveness, emotional state, likes/dislikes, or anything they felt would be a part of their first impression of the person they were seeing. Participants were given additional information in the instructions, depending on what information group they were assigned to. Those in correct and incorrect information groups were told the faces would be labelled as AI-generated or Real, while those in the no information group were given no additional information. The faces remained on screen until participants progressed to the next page. An inter-stimulus fractal noise mask was presented for 1 second between each trial. Images were presented at a size of 375 by 375 pixels, which, on the computer in our lab, subtended 10.6 by 10.6 degrees at a viewing distance of 57.3cm.

Once the experiment was complete, participants were debriefed regarding the information condition manipulation. They were then asked whether they believed the manipulation that they were exposed to and were given the option to have their data excluded from the study. Importantly, most participants indicated that they believed the manipulation (n = 69, 85%).

Moreover, several participants that indicated they did not believe the manipulation indicated that it was because in both image groupings (real and AI-generated), they felt a few did not seem to match the label given, rather than thinking they were being given completely incorrect information. They were then asked about their familiarity with and opinion on AI-generation and if they were distracted during the task. Finally, they were asked to provide any general comments on the study.

#### *4.3.4. Coding Scheme*

Data has been made available here: https://osf.io/gpu4d/overview?view_only=6870245fe32645cc821c28ebf47dde03. To analyse the participants' open-ended responses, a coding scheme was developed prior to data analysis and refined throughout data coding (see below for details). The final scheme incorporated elements of several models: 1) For valence and arousal, Feldman and Russell's (1998) semantic structure of affect was followed, 2) For warmth and dominance, the Revised Interpersonal Adjectives Scale (IAS-R; Gurtman & Pincus, 2000) was used, and 3) Elements of the Interpersonal Circumplex (IPIP-IPC; Markey & Markey, 2009) model were also employed. Using these models as guides, the authors then further refined the coding scheme after seeing common terms participants were using the describe the images. For example, for warmth, anything denoting agreeableness or approachability was considered warm (friendly, kind, polite, concern for another person, etc.). On the opposite side of the scale, terms that showed coldness or rejection were coded as such (anger, frustration, meanness, disgust, not wanting to talk/be approached). For dominance, terms that indicated control over another person, needing to be the centre of attention, or more dominant emotions such as anger we considered high dominance. In contrast,

terms that denoted not having control, meekness, stress, nervousness, concern, sweetness, etc. were considered low dominance.

Importantly, the coding scheme incorporates positive and negative valence, arousal, warmth, and dominance. However, the faces represented only two emotional expressions: anger and happiness. Because of this, it is important to note when interpreting the results that negative valence does not necessarily mean anger. Rather, several emotions and states were considered negatively valenced (e.g., anger, sadness, disgust, rudeness, meanness, etc.). Similarly for positive valence, this does not represent happiness, but rather positive emotions or states (e.g., happiness, excitement, calmness, relaxation, etc.). For arousal, it is important to remember that valence (although it can be correlated) is not directly related to arousal level. Both positively and negatively valenced faces were considered high and low arousal. Some terms used for negative arousal were calm, relaxed, bored, tired, sad, etc. Some terms for positive arousal were excited, happy, nervous, angry, etc.

#### *4.3.5. Data Analysis*

Prior to any coding or analysis, the data was cleaned. To reiterate, participants were first rejected based on incorrect responses to catch trials, and outlying duration. This was done for the entire dataset. Further data cleaning was done after data coding (see below).

Responses were then extracted from the data so that the coders would not see what knowledge group participants were assigned to.

Two coders coded the data separately. The coders were two of the authors who both identified as young adult women, one coder identified as White, and one coder identified as South Asian. Coders met after coding 5 participants to resolve any discrepancies with the coding

scheme. Coders then met approximately halfway through coding for each of the variables to again refine the coding scheme. Coders then separately coded the entire set of data. The scheme used was as follows: coders read participant responses and rated them as either negative (-1), neutral (0), or positive (+1); and as either low (-1), neutral/not enough information (0), or high (+1) on aspects of the Interpersonal Circumplex Model. Namely, they were coded on emotional valence, emotional arousal, warmth, and dominance. Several interpretations of the model were used to aid in our coding (Feldman Barrett & Russell, 1998; Gurtman & Pincus, 2000; Markey & Markey, 2009; Wiggins et al., 1988). It should be note that we originally planned to use a coding scheme involving the BIG 5 personality traits, using the TIPI (Gosling et al., 2003) as a guideline. However, after an initial examination of the raw data we determined that this would not provide a valid and reliable way to code it.

Following coding and meeting to resolve discrepancies the coders reached a minimum of 90% agreement on each dependent variable (DV, range: 90% - 99% agreement). Since agreement did not reach 100%, the two coders' ratings were averaged. Ratings were then averaged across trials within each condition. For example, emotional valence ratings for the three AI-generated, happy, Black men's faces were averaged, resulting in one value for emotional valence of AI-generated, happy, Black men. This resulted in 24 average ratings for each DV. The data was then cleaned based on these average ratings for each DV separately. Again, using the VanSelst and Jolicoeur (1994) guidelines, participants either had their score Winsorized (if 3 or fewer scores were outlying) or participants were removed entirely (if they had more than 3 outlying scores).

The final sample sizes for each DV are as follows: (1) Emotional Valence N = 79, (2) Emotional Arousal N = 81, (3) Warmth N = 81, (4) Dominance N = 80. Because in all cases we

did not reach our intended sample size, sensitivity analyses were run to determine the smallest effect size we could detect. Using equation 4c reversed in Goulet and Cousineau (2019), our actual adjusted sample size was used to calculate the sample size in G-power for each of our adjusted samples. A sensitivity analysis was run for each separately for F-test mixed measures design, independent groups effect, power of .95. The actual effect sizes we could detect ranged from $f = .2502$ to .2537.

Following this, several 2 × 2 × 2 × 3 × 3 mixed measures ANOVAs were run with Image Type (real and AI-generated), Stimulus Gender (man and woman), Emotional Expression (angry and happy), and Stimulus Race (Asian, Black, and White) as repeated measures variables, and Knowledge Group (correct information, incorrect information, and no information) as an independent groups variable. The dependent variables were coded emotional valence, emotional arousal, warmth, and dominance. In any case where the assumption of sphericity was violated, adjusted Greenhouse-Geisser values were used. Box's M was significant for all DVs and therefore we used a stricter criterion to detect interactions with the Information Group ($p \leq .001$). Only the highest level significant interactions are reported, for lower level interactions and main effects see Supplemental Materials Table S2-S5. For all pairwise comparisons, Bonferroni adjusted values are reported.

## 4.4. Results

### *4.4.1. Emotional Valence*

The highest level interaction revealed by the omnibus ANOVA was a significant 4-way interaction between the repeated measures variables [$F(2, 152) = 9.49$, $p < .001$, $\eta^2 = .11$].

This interaction was broken down by Emotional Expression, and two 3-way ANOVAs were run with Stimulus Type, Stimulus Gender, and Stimulus Race as independent variables; one for angry faces and one for happy faces.

#### 4.4.1.1. Angry Faces.

Results are illustrated in Figure 1. Within angry faces, the highest level interaction was the significant 3-way interaction [$F(1.77, 137.94) = 39.58$, $p < .001$, $\eta^2 = .34$]. This was then broken down by Stimulus Gender, where two 2-way ANOVAs were run with Image Type and Stimulus Race as dependent variable, one for men's faces and one for women's faces.

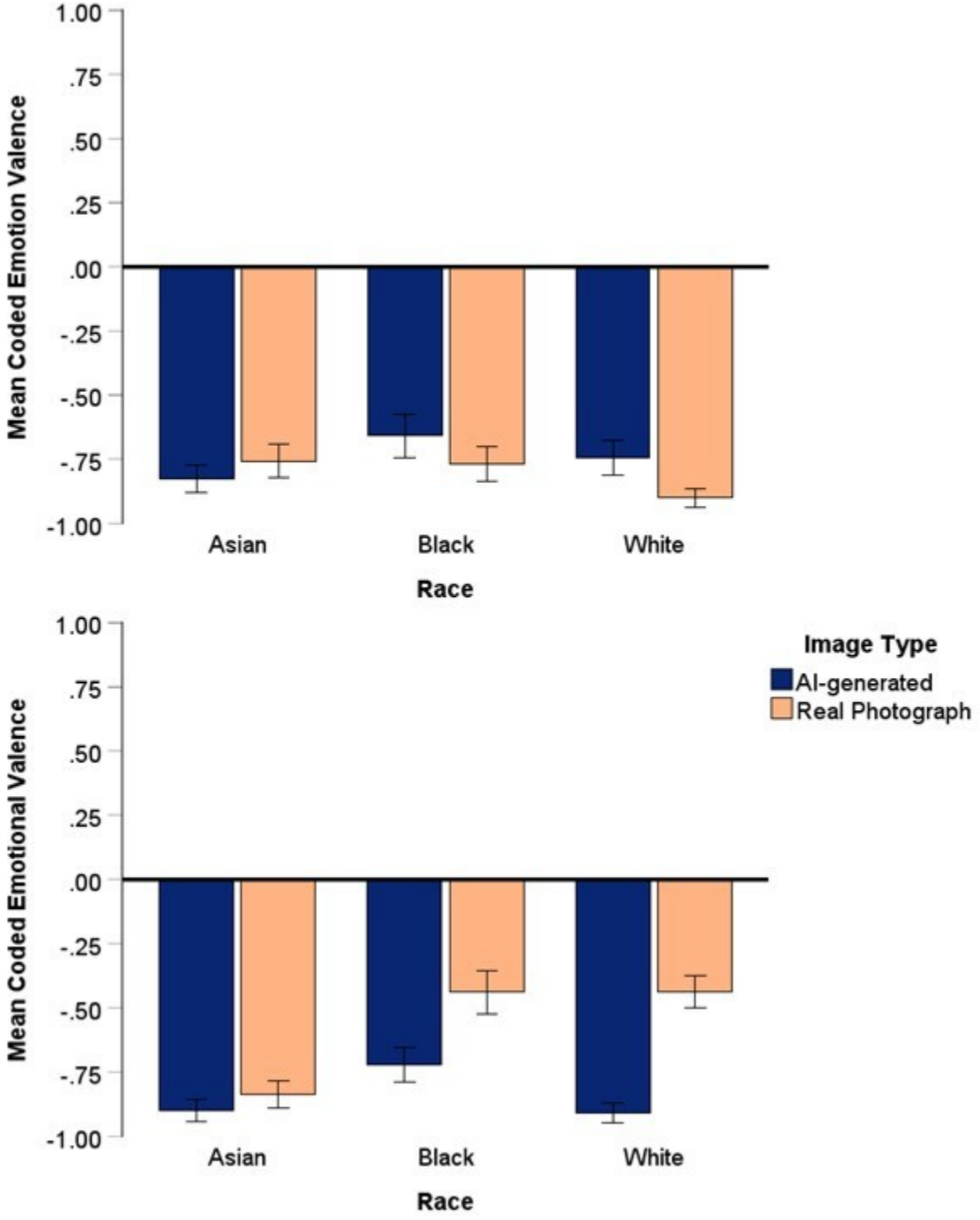
The highest level significant interaction for men's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 156) = 11.03$, $p < .001$, $\eta^2 = .12$]. Pairwise comparisons revealed that for Black and White faces, AIFs were significantly less negatively valenced than RFs [Black: $t(78) = 2.29$, $p = .026$, $d = .26$; White: $t(78) = 4.16$, $p < .001$, $d = .47$]. It also revealed that for AIFs, Black faces were significant less negatively valenced than Asian faces [$t(78) = 3.84$, $p < .001$, $d = .43$]. For RFs, Asian and Black faces were significantly less negatively valenced than White faces, with no difference between the two [Asian vs. White: $t(78) = 4.21$, $p < .001$, $d = .47$; Black vs. White: $t(78) = 4.29$, $p < .001$, $d = .48$].

The highest level significant interaction for women's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(1.82, 142.19) = 27.93$, $p < .001$, $\eta^2 = .26$]. Pairwise comparisons revealed that for all races, AIFs were more negatively valenced compared to RFs [Asian: $t(78) = 2.10$, $p = .04$, $d = .24$; Black: $t(78) = 5.55$, $p < .001$, $d = .62$; White: $t(78) = 12.54$, $p < .001$, $d = 1.40$]. Moreover, within AIFs, Black faces were less negatively valenced than both Asian and White faces [Black vs. Asian: $t(78) = 4.66$, $p < .001$, $d = .52$; Black vs.

White: $t(78) = 5.08$, $p < .001$, $d = .57$]. This pattern was not the same for RFs where Asian faces were more negatively valenced than Black and White faces [Asian vs. Black: $t(78) = 8.87$, $p < .001$, $d = 1.00$; Asian vs. White: $t(78) = 10.28$, $p < .001$, $d = 1.16$].

**Figure 1.**

*Mean Coded Emotional Valence for Angry Men and Women's Faces*



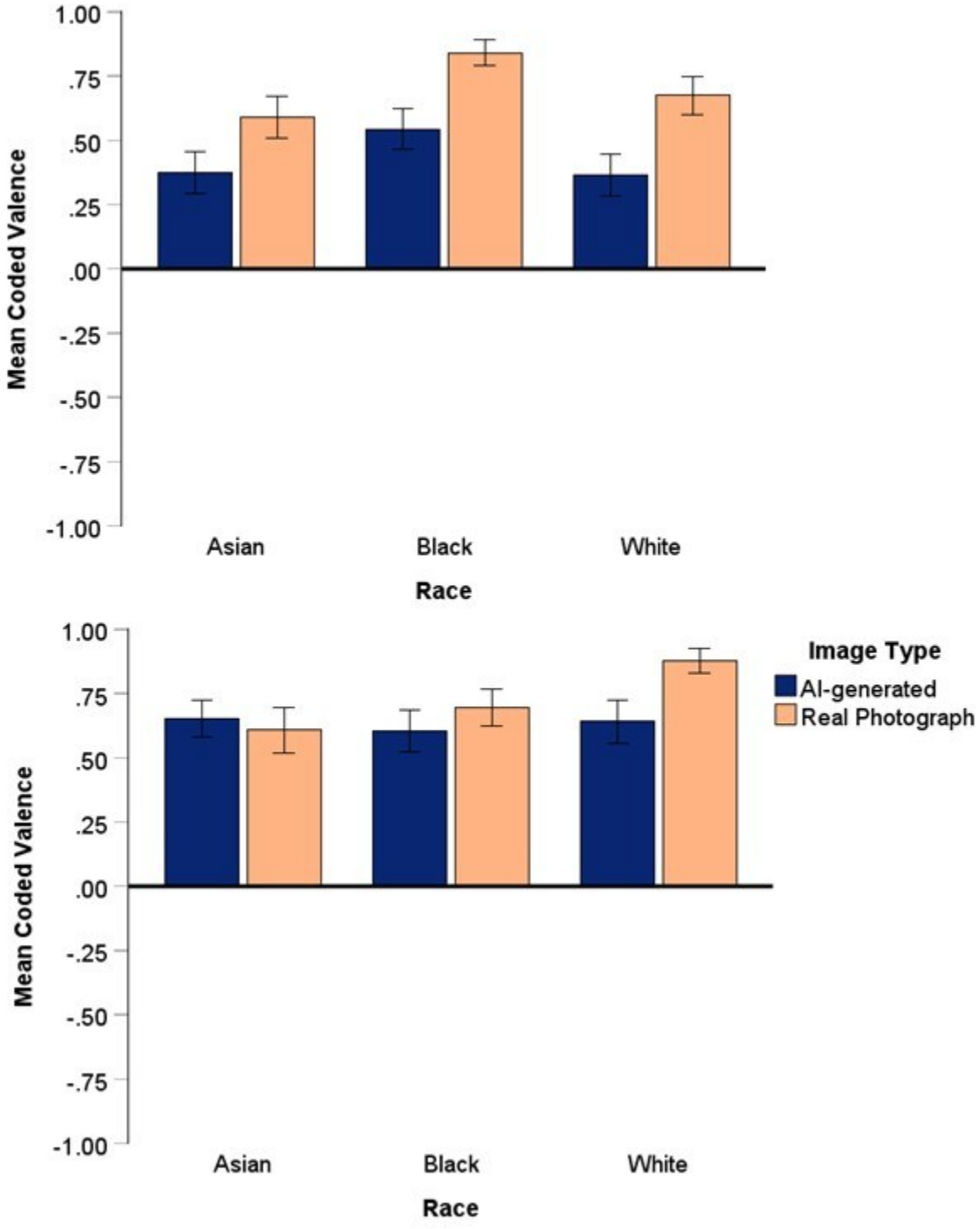

*Note.* Graphs represent the 2-way interactions between Image Type and Stimulus Race for angry faces for men (top) and women (bottom). The coding scale used was +1 (positive valence), 0 (neutral-N/A), -1 (negative valence). Error bars represent 95% CI.

#### 4.4.1.2. Happy Faces.

Results are illustrated in Figure 2. Within happy faces, the highest level interactions were the 2-way interactions between Image Type and Gender [$F(1, 78) = 26.93$, $p < .001$, $\eta^2 = .26$], Image Type and Race [$F(2, 156) = 11.01$, $p < .001$, $\eta^2 = .12$], and Gender and Race [$F(1.83, 143.01) = 29.79$, $p < .001$, $\eta^2 = .28$]. For both interactions involving Image type, in all cases AIFs were less positively valenced compared to RFs ($t$-values ranging from 2.13 - 8.50, $p$-values ranging from <.001 - .04, $d$-values ranging from .24 - .96). For the interaction between Gender and Race, pairwise comparisons revealed that for Asian and White faces, women's faces were more positively valenced, whereas for Black faces there was no difference between men and women [Asian: $t(78) = 4.97$, $p < .001$, $d = .56$; White: $t(78) = 8.81$, $p < .001$, $d = .99$].

**Figure 2.**

*Mean Coded Emotional Valence for Happy Men and Women's Faces*




*Note.* Graphs represent 2-way interactions between Image Type and Stimulus Race for happy faces for men (top) and women (bottom). The coding scale used was +1 (positive valence), 0 (neutral-N/A), -1 (negative valence). Error bars represent 95% CI.

### *4.4.2. Emotional Arousal*

The highest level interaction revealed by the omnibus ANOVA was a significant 4-way interaction between the repeated measures variables [$F(2, 156) = 14.58$, $p < .001$, $\eta^2 = .16$].

This interaction was broken down by Emotional Expression, and two 3-way ANOVAs were run with Stimulus Type, Stimulus Gender, and Stimulus Race as independent variables; one for angry faces and one for happy faces.

#### 4.4.2.1. Angry Faces.

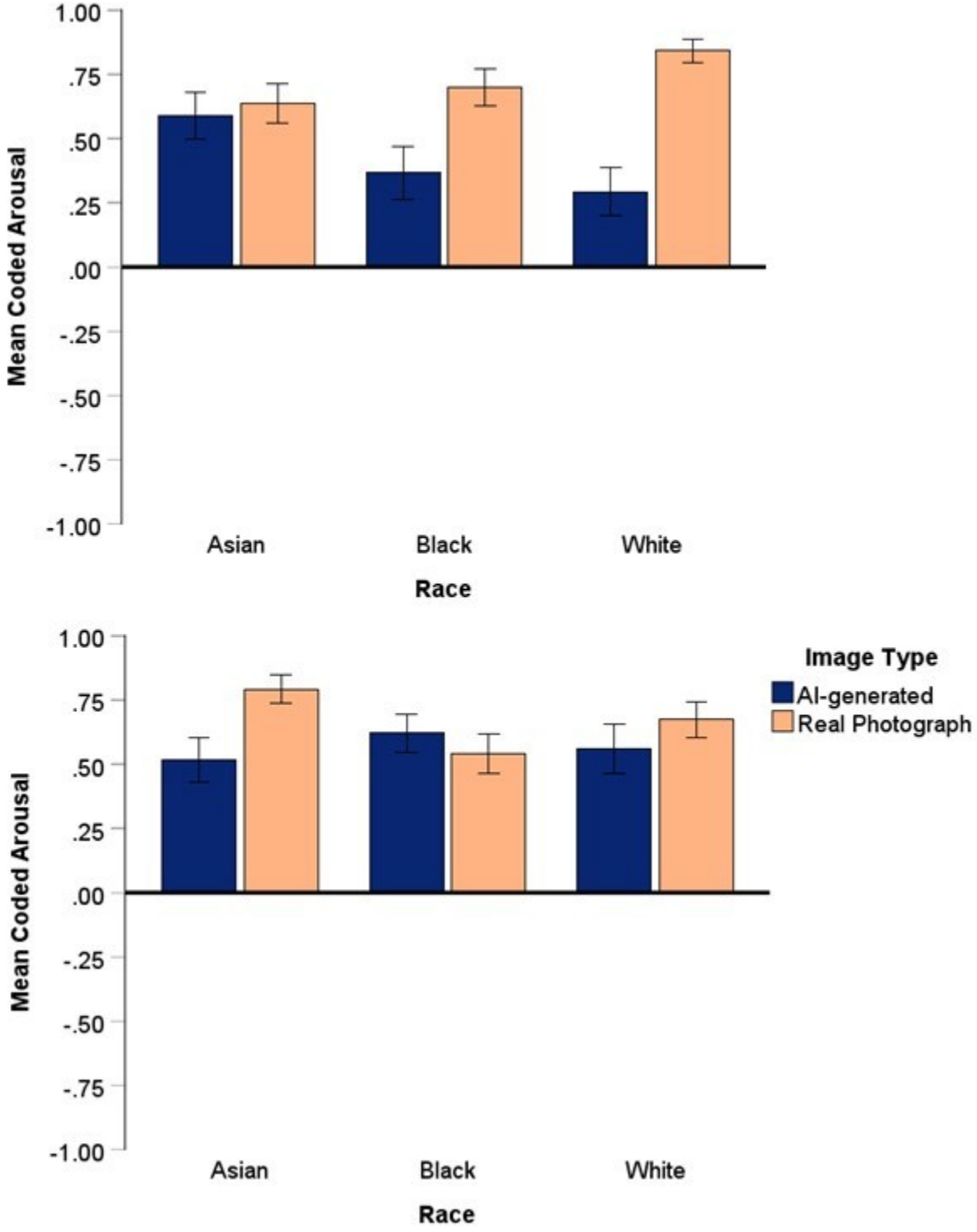
Results are illustrated in Figure 3. Within angry faces, the highest level interaction was the significant 3-way interaction [$F(2, 160) = 27.90$, $p < .001$, $\eta^2 = .26$]. This was then broken down by Stimulus Gender, where two 2-way ANOVAs were run with Image Type and Stimulus Race as dependent variable; one for men's faces and one for women's faces.

The highest level significant interaction for men's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 160) = 34.13$, $p < .001$, $\eta^2 = .30$]. Pairwise comparisons revealed that for Black and White faces, AIFs were significantly less arousing (though still positively arousing) than RFs, but no difference for Asian faces [Black: $t(80) = 5.98$, $p < .001$, $d = .66$; White: $t(80) = 10.17$, $p < .001$, $d = 1.13$]. It also revealed that the pattern across races was different for AIFs compared to RFs. For AIFs, Asian faces were more arousing than both Black and White faces, with no difference between Black and White faces [Asian vs. Black: $t(80) = 4.27$, $p < .001$, $d = .47$; Asian vs. White: $t(80) = 5.92$, $p < .001$, $d = .66$]. For RFs, White faces were significantly more arousing than Asian and Black faces, with no difference between Asian and Black faces [White vs. Asian: $t(80) = 5.23$, $p < .001$, $d = .58$; White vs. Black: $t(80) = 4.12$, $p < .001$, $d = .46$].

The highest level significant interaction for women's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 160) = 12.46$, $p < .001$, $\eta^2 = .14$]. Pairwise comparisons revealed that for Asian and White faces, AIFs were significant less arousing than their RF counterparts [Asian: $t(80) = 5.21$, $p < .001$, $d = .58$; White: $t(80) = 2.09$, $p = .04$, $d = .23$]. Moreover, for AIFs, there were no differences between races in terms of arousal, whereas for RFs, Asian faces were significantly more arousing than both Black and White faces, and White faces were significantly more arousing than Black faces [Asian vs. Black: $t(80) = 5.70$, $p < .001$, $d = .63$; Asian vs. White: $t(80) = 3.55$, $p = .002$, $d = .39$; Black vs. White: $t(80) = 3.12$, $p = .009$, $d = .35$].

**Figure 3.**

*Mean Coded Arousal for Angry Men and Women's Faces*




*Note.* Graphs represent 2-way interactions between Image Type and Race for angry faces of men (top) and women (bottom). The coding scale used was +1 (high arousal), 0 (neutral-N/A), -1 (low arousal). Error bars represent 95% CI.

#### 4.4.2.2. Happy Faces.

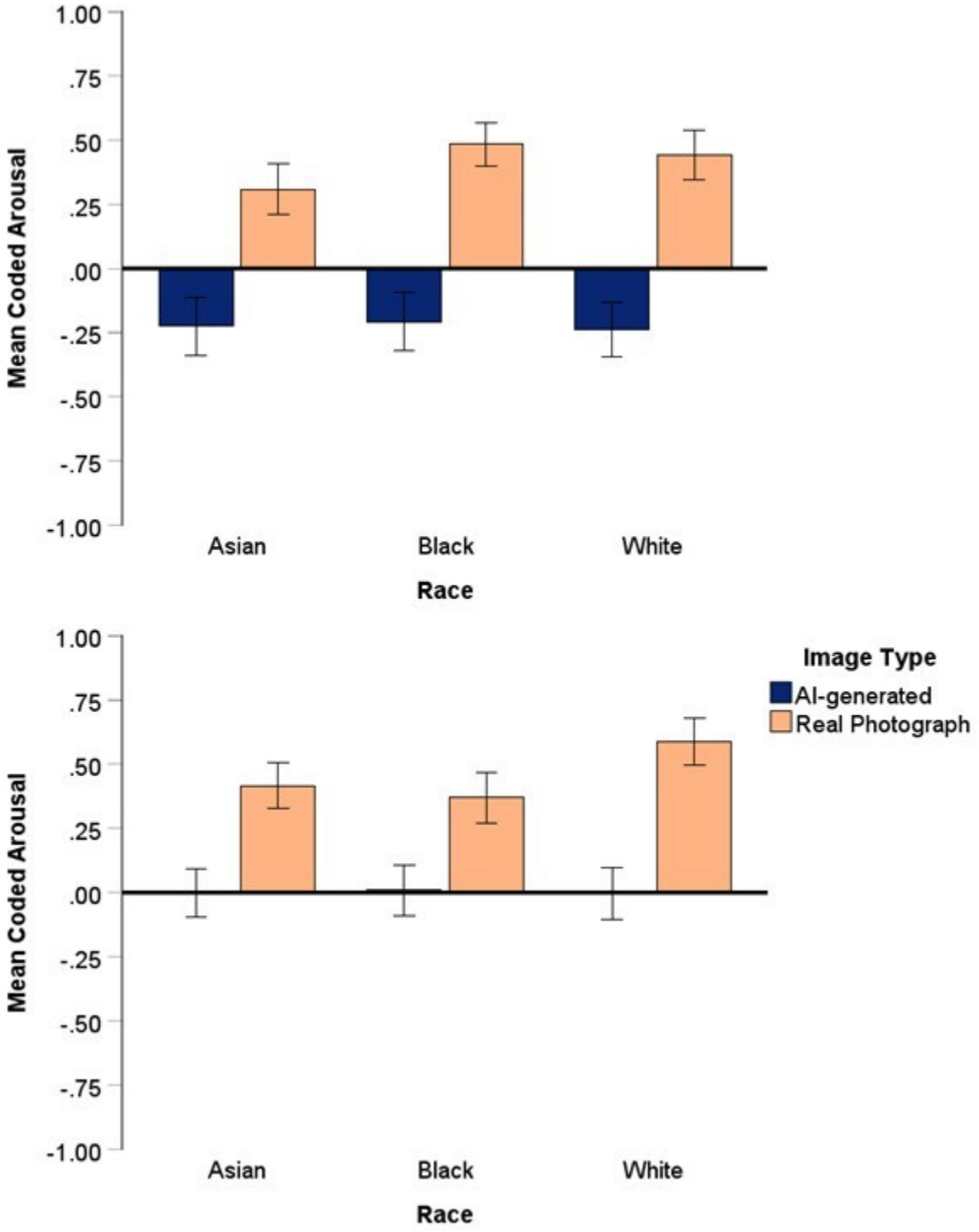
Results are illustrated in Figure 4. Within angry faces, the highest level interaction was the significant 3-way interaction [$F(2, 160) = 3.58$, $p = .03$, $\eta^2 = .04$]. This was then broken down by Stimulus Gender, where two 2-way ANOVAs were run with Image Type and Stimulus Race as dependent variable, one for men's faces and one for women's faces.

The highest level significant interaction for men's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 160) = 3.08$, $p = .05$, $\eta^2 = .04$]. Pairwise comparisons revealed that for all races, AIFs were significantly less arousing (and went into negative arousal in all cases) than RFs [Asian: $t(80) = 7.51$, $p < .001$, $d = .83$; Black: $t(80) = 10.63$, $p < .001$, $d = 1.18$; White: $t(80) = 10.32$, $p < .001$, $d = 1.15$]. It also revealed that the pattern across races was different for AIFs compared to RFs. For AIFs, there were no significant differences between races. For RFs, Asian faces were significantly less arousing than both Black and White faces, with no difference between Black and White faces [Asian vs. Black: $t(80) = 3.43$, $p = .003$, $d = .38$; Asian vs. White: $t(80) = 2.58$, $p = .04$, $d = .29$].

The highest level significant interaction for women's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 160) = 5.67$, $p = .004$, $\eta^2 = .07$]. Pairwise comparisons revealed that for all races, AIFs were significantly less arousing than their RF counterparts [Asian: $t(80) = 6.97$, $p < .001$, $d = .77$; Black: $t(80) = 5.93$, $p < .001$, $d = .66$; White: $t(80) = 8.85$, $p = .04$, $d = .98$]. Moreover, for AIFs, there were no differences between races in terms of arousal. For RFs, White faces were significantly more arousing than both Asian and Black faces, with no difference between Asian and Black faces [White vs. Asian: $t(80) = 4.22$, $p < .001$, $d = .47$; White vs. Black: $t(80) = 4.04$, $p < .001$, $d = .45$].

**Figure 4.**

*Mean Coded Arousal for Happy Men and Women's Faces*




*Note.* Graphs represent 2-way interactions between Image Type and Race for happy faces of men (top) and women (bottom). The coding scale used was +1 (high arousal), 0 (neutral-N/A), -1 (low arousal). Error bars represent 95% CI.

### *4.4.3. Warmth*

The highest level interaction revealed by the omnibus ANOVA was a significant 4-way interaction between the repeated measures variables [$F(2, 156) = 3.42$, $p = .04$, $\eta^2 = .04$].

This interaction was broken down by Emotional Expression, and two 3-way ANOVAs were run with Stimulus Type, Stimulus Gender, and Stimulus Race as independent variables; one for angry faces and one for happy faces.

#### 4.4.3.1. Angry Faces.

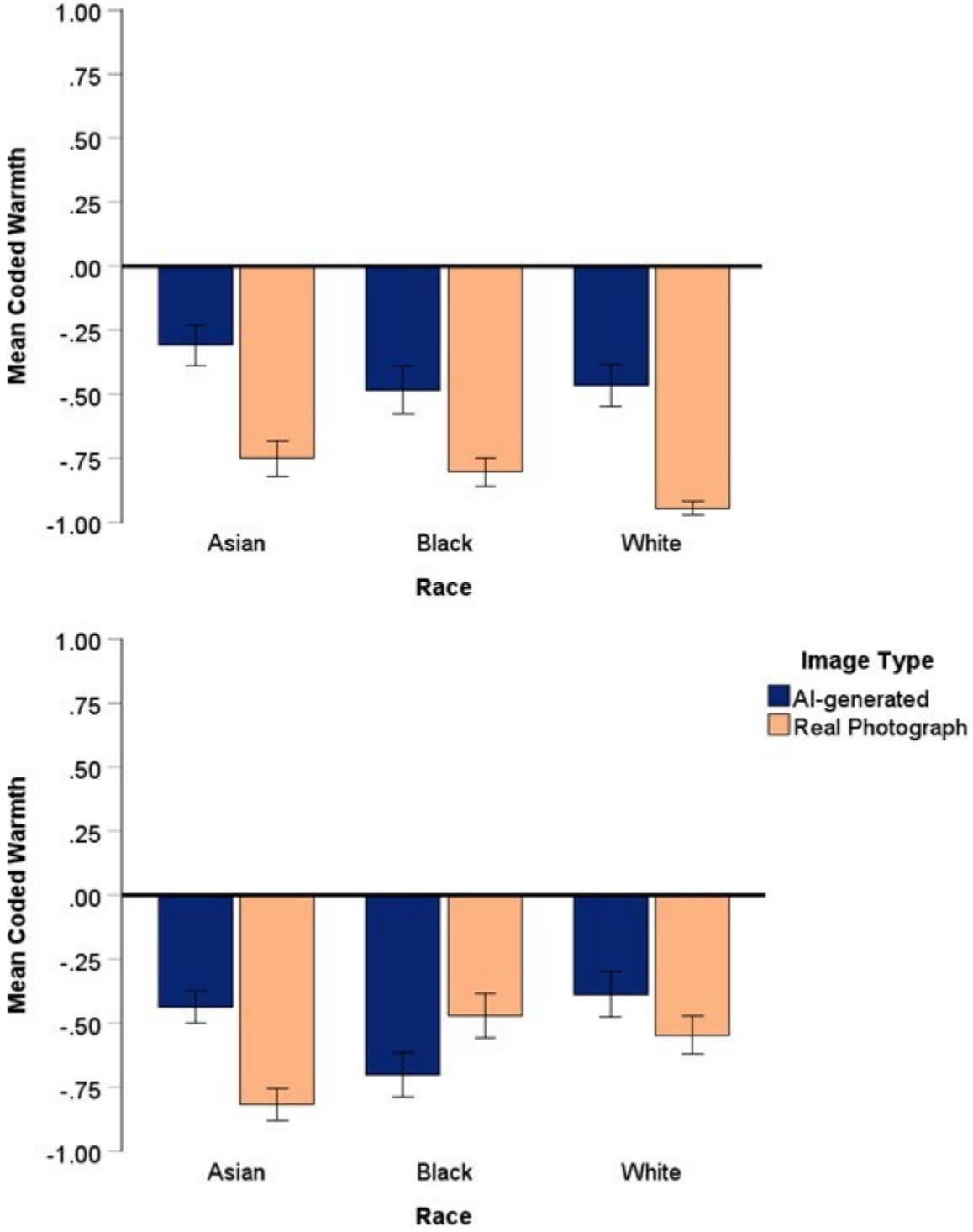
Results are illustrated in Figure 5. Within angry faces, the highest level interaction was the significant 3-way interaction [$F(1.81, 144.58) = 16.65$, $p < .001$, $\eta^2 = .17$]. This was then broken down by Stimulus Gender, where two 2-way ANOVAs were run with Image Type and Stimulus Race as dependent variable, one for men's faces and one for women's faces.

The highest level significant interaction for men's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 160) = 4.96$, $p = .008$, $\eta^2 = .06$]. Pairwise comparisons revealed that for all races, AIFs were significantly more warm (though still on the cold side of the scale) than RFs [Asian: $t(80) = 10.52$, $p < .001$, $d = 1.17$; Black: $t(80) = 6.42$, $p < .001$, $d = .71$; White: $t(80) = 11.19$, $p < .001$, $d = 1.24$]. It also revealed that the pattern across races was different for AIFs compared to RFs. For AIFs, Asian faces were warmer (though still considered cold) than both Black and White faces, with no difference between Black and White faces [Asian vs. Black: $t(80) = 3.71$, $p = .001$, $d = .41$; Asian vs. White: $t(80) = 3.12$, $p = .008$, $d = .35$]. For RFs, White faces were significantly more cold than Asian and Black faces, with no difference between Asian and Black faces [White vs. Asian: $t(80) = 5.13$, $p < .001$, $d = .57$; White vs. Black: $t(80) = 4.90$, $p < .001$, $d = .54$].

The highest level significant interaction for women's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(1.87, 149.34) = 45.32$, $p < .001$, $\eta^2 = .36$]. Pairwise comparisons revealed that for Asian and White faces, AIFs were significantly warmer (though still on the cold side of the scale) than their RF counterparts [Asian: $t(80) = 10.30$, $p < .001$, $d = .1.14$; White: $t(80) = 2.71$, $p = .008$, $d = .30$]. Whereas Black faces were significantly colder than their RF counterparts [$t(80) = 4.48$, $p < .001$, $d = .50$]. Moreover, for AIFs, Black faces were significantly colder than Asian and White faces [Black vs. Asian: $t(80) = 5.84$, $p <.001$, $d = .65$; Black vs. White: $t(80) = 5.83$, $p < .001$, $d = .65$]. However, for RFs, Asian faces were significantly colder than both Black and White faces, with no difference between Black and White faces [Asian vs. Black: $t(80) = 7.29$, $p < .001$, $d = .81$; Asian vs. White: $t(80) = 6.04$, $p < .001$, $d = .67$].

**Figure 5.**

*Mean Coded Warmth for Angry Men and Women's Faces*




*Note.* Graphs represent 2-way interactions between Image Type and Race for angry faces of men (top) and women (bottom). The coding scale used was +1 (warm), 0 (neutral-N/A), -1 (cold). Error bars represent 95% CI.

#### 4.4.3.2. Happy Faces.

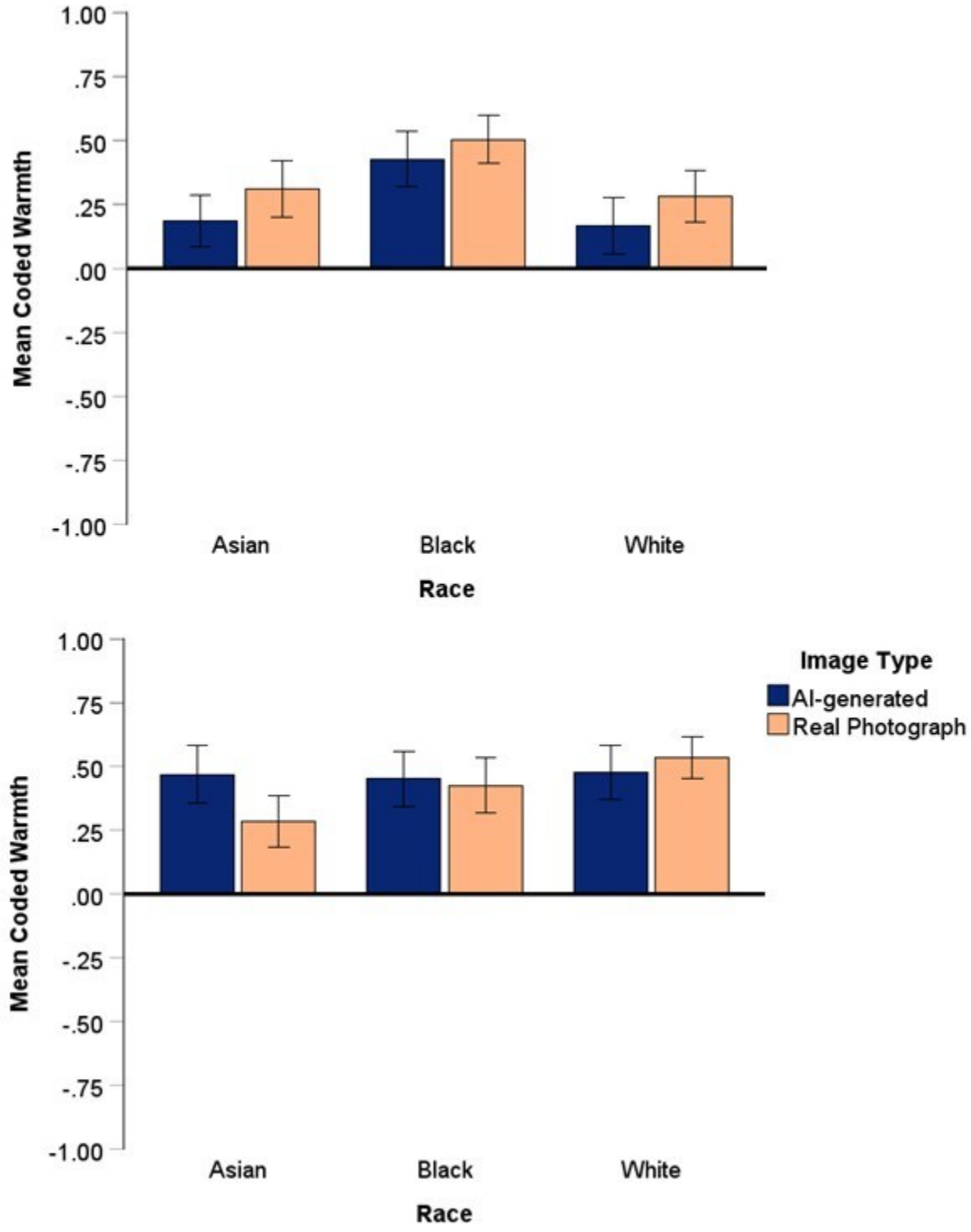
Results are illustrated in Figure 6. Within happy faces, the highest level interaction was the significant 3-way interaction [$F(2, 160) = 4.00$, $p = .02$, $\eta^2 = .05$]. This was then broken down by Stimulus Gender, where two 2-way ANOVAs were run with Image Type and Stimulus Race as dependent variable, one for men's faces and one for women's faces.

The main effects of Image Type [$F(1, 80) = 7.14$, $p = .009$, $\eta^2 = .08$] and Race [$F(2, 160) = 18.56$, $p < .001$, $\eta^2 = .19$] were statistically significant. The interaction between the two was not statistically significant. AIFs were significantly colder (though still on the warm side of the scale) than RFs [$t(80) = 2.65$, $p = .009$, $d = .29$]. Black faces are significantly warmer than Asian and White faces, with no difference between Asian and White faces [Black vs. Asian: $t(80) = 4.84$, $p < .001$, $d = .54$; Black vs. White: $t(80) = 5.88$, $p < .001$, $d = .65$].

The highest level significant interaction for women's faces was the 2-way interaction between Stimulus Type and Stimulus Race [$F(2, 160) = 5.74$, $p = .004$, $\eta^2 = .07$]. Pairwise comparisons revealed that that only for Asian faces were AIFs significantly warmer than their RF counterparts [$t(80) = 2.73$, $p = .008$, $d = .30$]. There were no significant differences for Black and White faces. Moreover, there were no difference between races within AIFs. However, for RFs, Asian faces were significantly colder than both Black and White faces, and Black faces were significantly colder than White faces (though all were still on the warm side of the scale)[Asian vs. Black: $t(80) = 2.46$, $p = .05$, $d = .27$; Asian vs. White: $t(80) = 5.58$, $p < .001$, $d = .62$; Black vs. White: $t(80) = 2.58$, $p = .04$, $d = .29$].

**Figure 6.**

*Mean Coded Warmth for Happy Men and Women's Faces*




*Note.* Graphs represent 2-way interactions between Image Type and Race for happy men (top) and women (bottom). Coding scale used was +1 (warm), 0 (neutral-N/A), -1 (cold). Error bars represent 95% CI.

### *4.4.4. Dominance*

The highest level significant interactions revealed by the omnibus ANOVA were the 3-way interactions between Image Type, Gender and Race [$F(2, 154) = 15.65$, $p < .001$, $\eta^2 = .17$], Image Type, Gender, and Emotion [$F(1, 77) = 17.62$, $p < .001$, $\eta^2 = .19$], Image Type, Race, and Emotion [$F(2, 154) = 31.32$, $p < .001$, $\eta^2 = .29$], and Gender, Emotion, and Race [$F(1.78, 137.14) = 37.78$, $p < .001$, $\eta^2 = .33$]. To examine these effects further, they were broken down into smaller 2-way interactions and pairwise comparisons were analyzed.

#### 4.4.4.1. Image Type × Gender × Race Breakdown.

To examine the 3-way interaction between Image Type, Gender, and Race, the data was first collapsed across Emotion. Using this collapsed data, two 2-way ANOVAs were run between Image Type and Race, one for men [$F(2, 158) = 10.12$, $p < .001$, $\eta^2 = .11$] and one for women [$F(2, 158) = 53.81$, $p < .001$, $\eta^2 = .41$] and both were statistically significant. Results are illustrated in Figure 7.

##### *4.4.4.1.1. Image Type × Race: Men.*

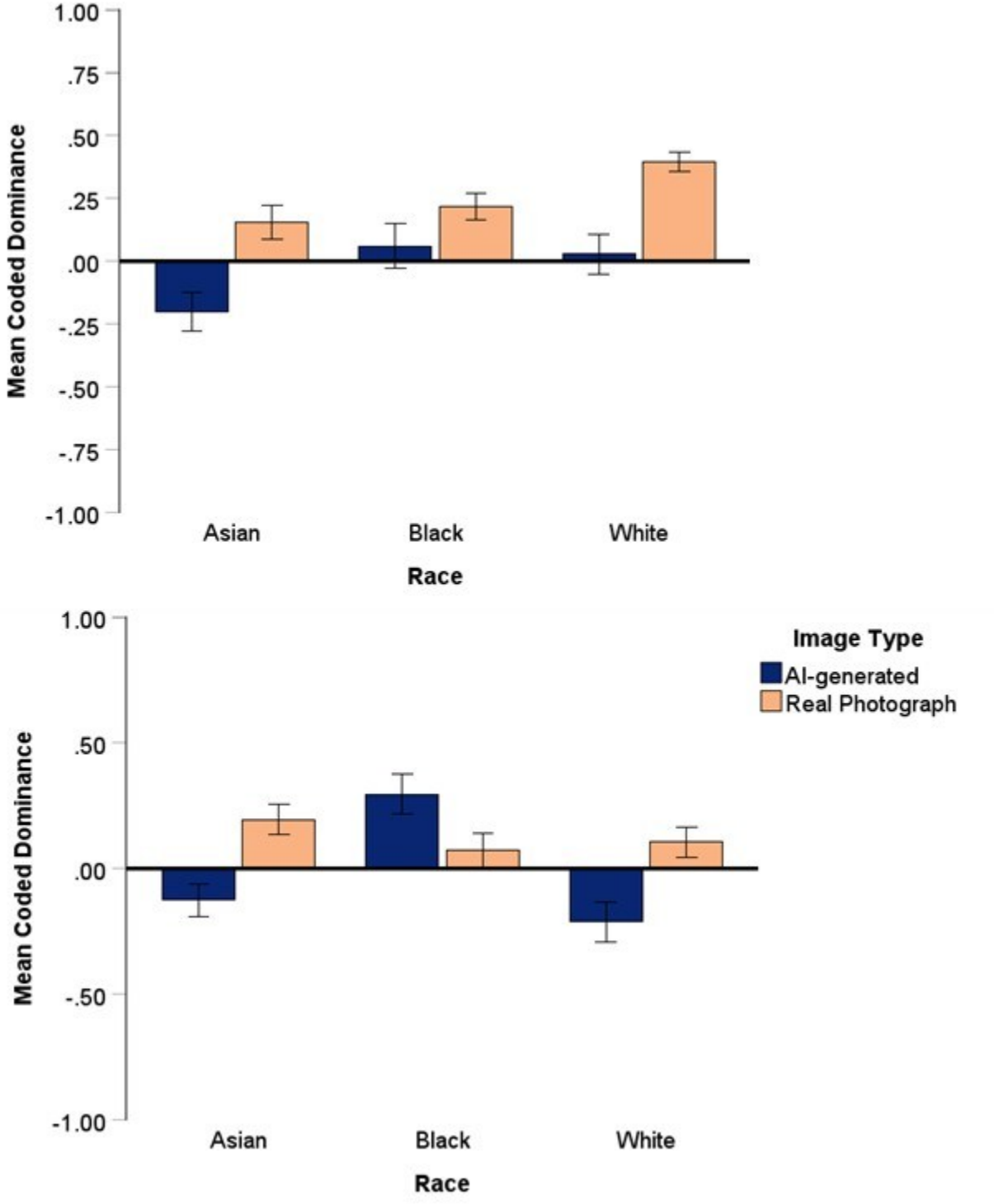
Pairwise comparisons revealed that, for men, within each race AIFs were less dominant than their RF counterparts [Asian: $t(79) = 7.78$, $p < .001$, $d = .87$; Black: $t(79) = 3.51$, $p < .001$, $d = .39$; White: $t(79) = 8.39$, $p < .001$, $d = .94$]. Moreover, within AIFs, Asian faces were significantly less dominant than both Black and White faces, with no difference between Black and White faces (Asian vs. Black: $t(79) = 5.80$, $p < .001$, $d = .65$; Asian vs. White: $t(79) = 5.75$, $p < .001$, $d = .64$]. Whereas for RFs, Asian and Black faces were significantly less dominant than White faces, with no difference between Asian and Black faces [White vs. Asian: $t(79) = 6.03$, $p < .001$, $d = .67$; White vs. Black: $t(79) = 6.63$, $p < .001$, $d = .74$].

#### *4.4.4.1.2. Image Type × Race: Women.*

Pairwise comparisons revealed a different pattern of results for women. For Asian and White faces AIFs were less dominant than their RF counterparts [Asian: $t(79) = 8.45$, $p < .001$, $d = .94$; White: $t(79) = 6.36$, $p < .001$, $d = .71$]. For Black faces, however, AIFs were *more* dominant than RFs [$t(79) = 4.33$, $p < .001$, $d = .48$]. Moreover, within AIFs, Asian and White faces were significantly less dominant than Black faces, with no difference between Asian and White faces [Black vs. Asian: $t(79) = 9.36$, $p < .001$, $d = 1.05$; Black vs. White: $t(79) = 10.37$, $p < .001$, $d = 1.16$]. However, for RFs, Asian faces were more dominant than Black faces [$t(79) = 3.29$, $p = .004$, $d = .37$]. No other differences were statistically significant.

**Figure 7.**

*Mean Coded Dominance Collapsed Across Emotion*




*Note.* Graphs represent 2-way interactions between Image Type and Race for men (top) and women (bottom). Coding scale used was +1 (high dominance), 0 (neutral-N/A), -1 (low dominance). Error bars represent 95% CI.

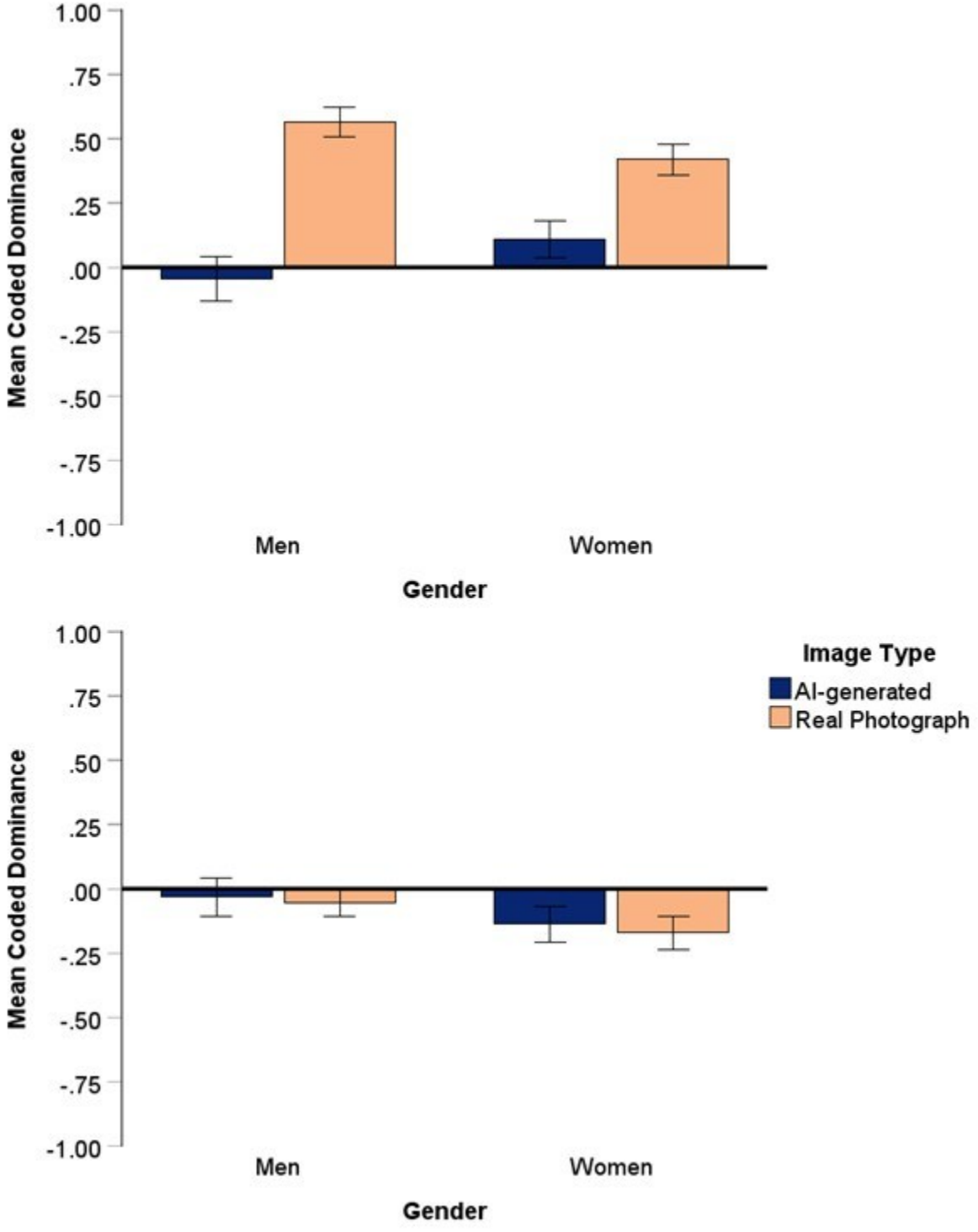
#### 4.4.4.2. Image Type × Gender × Emotion Breakdown.

To examine the 3-way interaction between Image Type, Gender, and Emotion, the data was first collapsed across Race. Using this collapsed data, two 2-way ANOVAs were run between Image Type and Gender, one for angry faces [$F(1, 79) = 29.76$, $p < .001$, $\eta^2 = .27$] and one for happy faces. Within happy faces, only the main effect of Gender was significant [$F(1, 79) = 19.84$, $p << .001$, $\eta^2 = .20$]. Results are illustrated in Figure 8.

Pairwise comparisons revealed that within angry faces, for both men and women, AIFs were significantly less dominant than RFs [men: $t(79) = 12$, $p < .001$, $d = 1.34$; women: $t(79) = 7.23$, $p < .001$, $d = .81$]. Moreover, the pattern was different when comparing men and women within AIFs and RFs. For AIFs, men were significantly less dominant than women [$t(79) = 3.56$, $p < .001$, $d = .40$], whereas for RFs, the opposite was true [$t(79) = 4.45$, $p < .001$, $d = .50$].

**Figure 8.**

*Mean Coded Dominance Collapsed Across Race*




*Note.* Graphs represent 2-way interactions between Image Type and Gender for angry faces (top) and happy faces (bottom). Coding scale used was +1 (high dominance), 0 (neutral-N/A), -1 (low dominance). Error bars represent 95% CI.

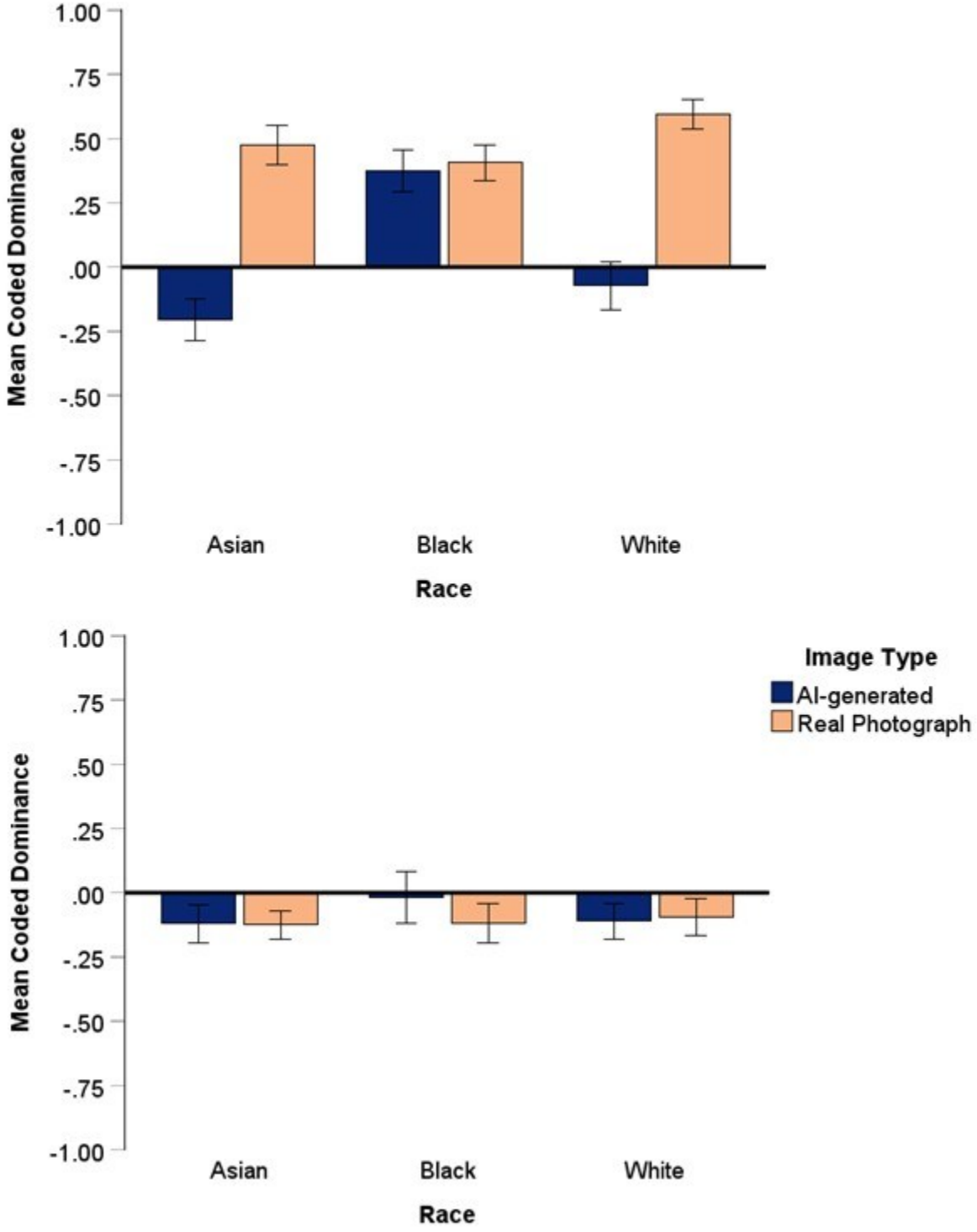
#### 4.4.4.3. Image Type × Race × Emotion Breakdown.

To examine the 3-way interaction between Image Type, Race and Emotion, the data was first collapsed across Gender. Using this collapsed data, two 2-way ANOVAs were run between Image Type and Race, one for angry faces [$F(2, 158) = 90.59$, $p < .001$, $\eta^2 = .53$] which was statistically significant, and one for happy faces (neither main effect nor the interaction were significant). Results are illustrated in Figure 9.

Pairwise comparisons revealed that within angry faces, Asian and White AIFs were less dominant than their RF counterparts, whereas there was no difference for Black faces [Asian: $t(79) = 14.53$, $p < .001$, $d = 1.62$; White: $t(79) = 12.39$, $p < .001$, $d = 1.39$]. Moreover, the pattern for AIFs was different from RFs across races. For AIFs, Black faces were significantly more dominant than Asian and White faces (which were both on the low side of the scale), and White faces were significantly more dominant than Asian faces [Black vs. Asian: $t(79) = 12.61$, $p < .001$, $d = 1.41$; Black vs. White: $t(79) = 9.51$, $p < .001$, $d = 1.06$; White vs. Asian: $t(79) = 2.77$, $p = .02$, $d = .31$]. For RFs, White faces were significantly more dominant than both Asian and Black faces, with no difference between Asian and Black faces [White vs. Asian: $t(79) = 3.05$, $p = .01$, $d = .34$; White vs. Black: $t(79) = 4.42$, $p < .001$, $d = .49$].

**Figure 9.**

*Mean Coded Dominance Collapsed Across Gender*




*Note.* Graphs represent 2-way interaction between Image Type and Race for angry faces (top) and happy faces (bottom). Coding scale used was +1 (high dominance), 0 (neutral-N/A), -1 (low dominance). Error bars represent 95% CI.

#### 4.4.4.4. Gender × Emotion × Race Breakdown.

To examine the 3-way interaction between Gender, Race and Emotion, the data was first collapsed across Image Type. Using this collapsed data, two 2-way ANOVAs were run between Gender and Race, one for angry faces [$F(2, 158) = 64.82$, $p < .001$, $\eta^2 = .45$] and one for happy faces [$F(2, 158) = 18.68$, $p < .001$, $\eta^2 = .19$], both of which were statistically significant. Results are illustrated in Figure 10.

##### *4.4.4.4.1. Gender × Race: Angry Faces.*

Pairwise comparisons revealed for angry faces, Asian men's faces were significantly less dominant than Asian women's [$t(79) = 7.02$, $p < .001$, $d = .79$], whereas there was no difference between Black men and women's faces, and White men's faces were more dominant than White women's [$t(79) = 9.26$, $p < .001$, $d = 1.04$]. Moreover, the pattern was different for men and women across races. For men, Asian men's faces were significantly less dominant than both Black and White faces, with no difference between Black and White faces [Asian vs. Black: $t(79) = 7.29$, $p < .001$, $d = .89$; Asian vs. White: $t(79) = 9.93$, $p < .001$, $d = 1.11$]. Whereas for women's faces, Asian and White faces were significantly less dominant than Black faces, and White faces were significantly less dominant than Asian faces [Black vs. Asian: $t(79) = 2.87$, $p = .02$, $d = .32$; Black vs. White: $t(79) = 6.76$, $p < .001$, $d = .76$; Asian vs. White: $t(79) = 4.27$, $p < .001$, $d = .48$].

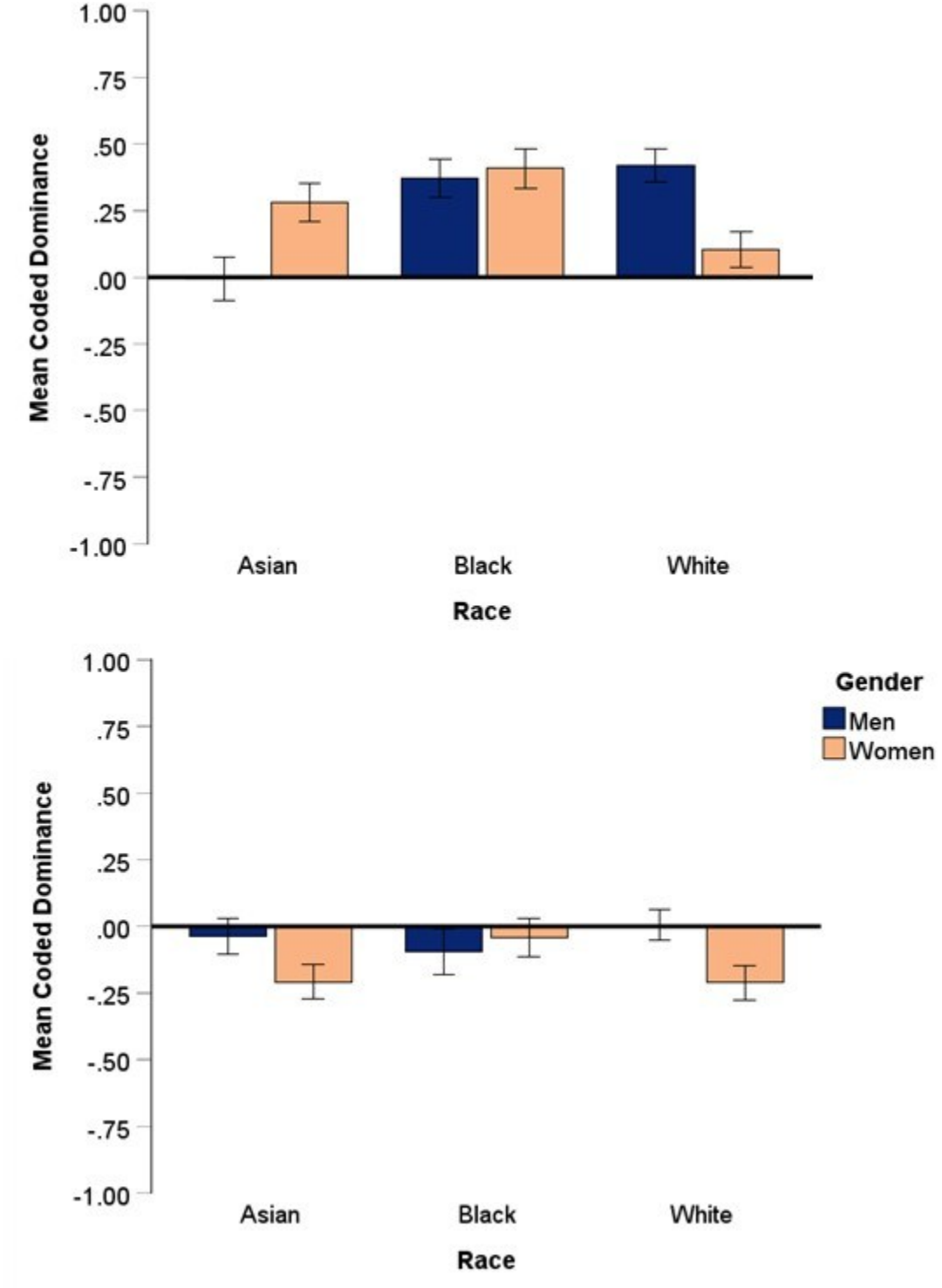
##### *4.4.4.4.2. Gender × Race: Happy Faces.*

Pairwise comparisons for happy faces revealed that, like angry faces, Asian men's faces were less dominant than Asian women's faces [$t(79) = 4.07$, $p < .001$, $d = .46$], whereas there was no difference between Black men and women's faces, and White men's faces were more dominant than White women's [$t(79) = 6.61$, $p < .001$, $d = .74$]. Moreover, the pattern was

different for men's and women's faces across races. For men's faces, the only significant difference was that White faces were significantly more dominant than Black faces [$t(79) = 2.73$, $p = .02$, $d = .31$]. Whereas for women's faces, Black faces were significantly less dominant than both Asian and White faces, with no difference between Asian and White faces [Black vs. Asian: $t(79) = 4.91$, $p < .001$, $d = .55$; Black vs. White: $t(79) = 5.31$, $p < .001$, $d = .59$].

**Figure 10.**

*Mean Coded Dominance Collapsed Across Image Type*




*Note.* Graphs represent 2-way interactions between Gender and Race for angry faces (top) and happy faces (bottom). Coding scale used was +1 (high dominance), 0 (neutral-N/A), -1 (low dominance). Error bars represent 95% CI.

### 4.5. Discussion

The discussion that follows will reiterate hypotheses and research questions and discuss how the results aid in our understanding of these concepts. Because all of the dependent variables, the effects of Gender, Race, and Emotional Expression did interact (often with Image Type), the discussion will focus on Q4.

***H1:*** *AI generators disproportionately create average and attractive faces unless otherwise specified. We therefore predict that knowledge about whether a face is AI-generated or real will impact the social attributions assigned to that face.*

Because of the use of mixed measures in this study, our sample size was slightly underpowered to detect differences between Information Groups. As with all null results, this should be interpreted with caution. Our sample size was sufficient to detect a medium effect size for this effect, but it is possible that there is an effect of knowledge that is smaller than we were able to detect. Recent literature has found that information about whether a face is AI-generated or a real photograph does impact how they are perceived (Liefooghe et al., 2023; Nightingale & Farid, 2022; Tucciarelli et al., 2022).

***Q4:*** *Will the factors of race, gender, and expression interact in determining the social attribution assigned to a face?*

The results support the notion that Race, Gender, and Emotional Expression interact (often with Image Type) to produce different perceptions of AIFs. Often, the pattern of results was different when comparing AIFs and RFs, when comparing men and women, and when comparing Asian, Black, and White faces. From the results, we see several stereotypes emerge; This is in line with previous work, which states that the training sets of AI-generators are biased,

and thus lead to reproduction and often amplification of societally held biases and stereotypes (AlDahoul et al., 2025; Assis & Moura, 2025; Bianchi et al., 2023; Chauhan et al., 2024; Gawali et al., 2024; Ghosh & Caliskan, 2023; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Mubashir, 2024 ; Park, 2024).

### *4.5.1. Stereotype Reproduction*

Several stereotypes emerged from our results. One of the most salient was the "angry black woman" trope (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009). This is a pervasive stereotype that persists despite a lack of evidence, by which Black women are depicted as angry, aggressive, and loud compared to others. The emergence of this stereotype in AI-generators is supported by the Emotional Valence and Warmth results, and to a lesser extent, Dominance. To reiterate, Emotional Valence was coded as negative, neutral, and positive (and negative valences could indicate anger, sadness, aggression, etc.), Warmth was coded as cold, neutral, and warm (and coldness could indicate aggression, rudeness, meanness, unapproachability, etc.), and Dominance was coded as low dominance (submissive), neutral, and high dominance (and high dominance could indicate aggression, threat, confidence, power, loudness, etc.). In the cases of Emotional Valence and Warmth, despite the RFs for Black women being the closest to neutral, the AIFs were significantly more negative and colder than the RFs, and in the case of Warmth, it was the most cold of any of the women's faces. In terms of Dominance, we see this stereotype emerge when collapsing across emotion (therefore this is interpreted with caution), where the AIFs of Black women are significantly more dominant than their RF counterparts and AIFs of Asian and White faces. Moreover, when collapsing across gender, we see that AIFs of Black faces have the same level of dominance as their RF counterparts, where for Asian and White faces the AIFs are

significantly less dominant than their RF counterparts. This last example could indicate an overall perception of Black faces (men or women) as more aggressive, loud, or hostile, which is in line with previous literature on stereotypes about Black individuals (Melson-Silimon et al., 2024). This is supported yet again when collapsing across image type, where we see that there is no difference between Black men and women's dominance, and it is often high (especially for anger) compared to the other two races. This should be interpreted with caution, however, as we don't see this consistently across the different dependent variables. Using Nvivo, we ran a word frequency query (minimum word length of four characters, grouped with stemmed words) and common words are as follows (in order of most prevalent to least prevalent): *angry, serious, confused, mean, and upset.*

Another stereotype that emerged from the results is that of the "passive/submissive/weak Asian woman (and often man)" (Azhar et al., 2021; Lee, 2013; Nguyen, 2016). This is supported in several areas where the AIFs for Asian women are consistently perceived as expressing lower intensity emotions than their RF counterparts and the stimuli portraying other races. For Asian men, we see this demonstrated in Warmth and Dominance. Specifically for Asian women, their angry faces were less negatively valenced, less arousing (happy faces also, but that was consistent across races), warmer, and less dominant. Moreover, their happy faces were considered warmer than RFs, where for Black and White faces, there was no difference (however, this is more likely due to a reduced warmth of the RFs).

For Asian men, we see this stereotype emerge with their angry faces being warmer than their RF counterparts, as well as the AIFs of both Black and White faces (however, although this difference was stronger in Asian men, Black and White men's faces did follow the same pattern).

Moreover, we see that their AIFs (regardless of emotion) are less dominant, but especially for anger.

Finally, we see that when collapsing across gender, angry, Asian AIFs overall are more submissive than their RF counterparts, and compared to Black and White faces, despite the RFs being relatively similar in dominance. Some examples of descriptors commonly applied to Asian women AIFs expressing anger after a word frequency query (minimum word length of four characters, grouped with stemmed words) are as follows (in order of most prevalent to least prevalent): *confused, angry, upset, worried, tired, and concerned.* Using the same parameters, examples of descriptors commonly applied to AIFs of Asian men expressing anger were as follows: *confused, worried, concerned, angry, stressed, upset, scared, and tired.* For both Asian men and women's faces, the terms used contrast meaningfully with those used for AIFs of Black women. The terms applied to Asian faces often suggest worry rather than outright anger, whereas those applies to AIFs of Black women more often imply threat or a desire to avoid the person.

#### *4.5.2. AI-generated vs. Real Face Photographs Boradly*

In cases other than the ones described above, where stereotypes are not emerging in the results, we often see an overall reduction in trait intensity for AIFs compared to RFs. We see this strongly demonstrated in the warmth and arousal of angry faces, the arousal of happy faces, and in certain instances of dominance. This supports the idea that emotions may not be accurately represented in the training sets of AI-generators (Lawrence, Cimermanis, & Collin, 2025; Lawrence & Collin, 2025).

### 4.6. Limitations and Future Work

To reiterate, this study was exploratory and slightly underpowered. Due to the use of mixed measures, the authors had to balance adequate power with feasibility of coding the

qualitative data. As such, the null findings regarding information group effects should be interpreted with caution, and future research should examine this using quantitative measures, where it is feasible to recruit more participants and enhance the power.

Another limitation is that we did not examine differences based on gender and race of the participants. As stated previously model *and observer* characteristics may influence social and emotional perceptions of faces (Campbell et al., 2010; Cortes et al., 2019; Kiiski et al., 2016; Olivola et al., 2014; Zebrowitz, 2017; Zhang et al., 2020). With a larger sample size, these effects could be examined and future research should aim to include participant race and gender as independent variables. Moreover, a more ethnically/racially diverse sample with equal groups should be recruited to aid in examining impacts of observer race/ethnicity.

Finally, because the data collected was qualitative and coded by two coders, future research should examine these effects using a quantitative approach. This will allow for recruitment of more participants and finer examination of the different dependent variables. Based on our findings, examining Emotional Valence, Arousal, Warmth, and Dominance using scales will enhance this area of research.

### 4.7. Conclusion

This exploratory study supports and adds to existing literature which posits that training sets of AI-generators are inherently biased, and thus reproduce and amplify existing social biases and stereotypes with regard to race and gender (AlDahoul et al., 2025; Assis & Moura, 2025; Bianchi et al., 2023; Chauhan et al., 2024; Gawali et al., 2024; Ghosh & Caliskan, 2023; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Mubashir, 2024 ; Park, 2024). Because use of AI-generators is becoming more common-place, this has implications for real world uses of AI, such as in facial recognition software, and more broadly, in our perceptions of others. As AI becomes

more realistic, the pervasiveness of stereotypes will contribute to the continuing belief that these stereotypes are true and inherent to the individuals most affected, despite lack of empirical evidence to support these notions.

**4.8. References**

AlDahoul, N., Rahwan, T., & Zaki, Y. (2025). *AI-generated faces influence gender stereotypes and racial homogenization* (arXiv:2402.01002). Sci Rep 15(1):14449. https://doi.org/10.1038/s41598-025-99623-3

Ames, D. R., & Bianchi, E. C. (2008). The Agreeableness Asymmetry in First Impressions: Perceivers' Impulse to (Mis)judge Agreeableness and How It Is Moderated by Power. *Personality and Social Psychology Bulletin*, *34*(12), 1719–1736. https://doi.org/10.1177/0146167208323932

Ashley, W. (2014). The angry black woman: The impact of pejorative stereotypes on psychotherapy with black women. *Social work in public health, 29*(1), 27-34. DOI: 10.1080/19371918.2011.619449

Assis, J. D., & Moura, M. A. (2025). Algorithmic Semiosis and Racial Bias: A Study of Images Created by Generative AI. *Encontros Bibli*, *30*, e103495. DOI: 10.5007/1518-2924.2025.e103495

Azhar, S., Alvarez, A. R., Farina, A. S., & Klumpner, S. (2021). "You're so exotic looking": An intersectional analysis of Asian American and Pacific Islander stereotypes. *Affilia, 36*(3), 282-301. https://doi.org/10.1177/08861099211001460

Bhardwaj, N., Bhardwaj, A., & Garg, L. (2025, February). Controlling Bias in Generative AI: Techniques for Fair and Equitable Data Generation in Socially Sensitive Applications. In *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1-8). IEEE.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. https://doi.org/10.1145/3593013.3594095

Campbell, D. W., Neuert, T., Friesen, K. B., & McKeen, N. A. (2010). Assessing Social Approachability: Individual Differences, In-Group Biases, and Experimental Control. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *42*(4), 254–263. https://doi.org/10.1037/a0020229

Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., & Vanga, R. (2024, February). Identifying race and gender bias in diffusion ai image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)* (pp. 1-6). IEEE. DOI: 10.1109/ICAIC60265.2024.10433840

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, *270*, 1059–1067. https://doi.org/10.1016/j.psychres.2018.04.066

Cortes, D. S., Link to external site, this link will open in a new window, Laukka, P., Link to external site, this link will open in a new window, Ebner, N. C., & Fischer, H. (2019). Age-related differences in evaluation of social attributes from computer-generated faces of varying intensity. *Psychology and Aging*, *34*(5), 686–697. https://doi.org/10.1037/pag0000364

Eiserbeck, A., Maier, M., Baum, J., & Abdel Rahman, R. (2023). Deepfake smiles matter less—The psychological and neural impact of presumed AI-generated faces. *Scientific Reports*, *13*(1), 16111. https://doi.org/10.1038/s41598-023-42802-x

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967–984. https://doi.org/10.1037/0022-3514.74.4.967

Gawali, A., Tarle, A., Can Inac, C., & Kulkarni, S. S. (2024). A Framework for Detecting Stereotypes, Prejudices and Discrimination in AI-Generated Imagery.

Ghosh, S., & Caliskan, A. (2023). 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6971–6985. https://doi.org/10.18653/v1/2023.findings-emnlp.465

Gorska, A. M., & Jemielniak, D. (2023). The invisible women: uncovering gender bias in AI-generated images of professionals. *Feminist Media Studies*, *23*(8), 4370-4375. https://doi.org/10.1080/14680777.2023.2263659

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Goulet, M. A., & Cousineau, D. (2019). The power of replicated measures to increase statistical power. *Advances in Methods and Practices in Psychological Science*, *2*(3), 199-213. https://doi.org/10.1177/2515245919849434

Gurtman, M. B., & Pincus, A. L. (2000). Interpersonal Adjective Scales: Confirmation of Circumplex Structure from Multiple Perspectives. *Personality and Social Psychology Bulletin*, *26*(3), 374–384. https://doi.org/10.1177/0146167200265009

Jääskeläinen, P., Sharma, N. K., Pallett, H., & Åsberg, C. (2025). Intersectional analysis of visual generative AI: the case of stable diffusion. *AI & SOCIETY*, 1-22. https://doi.org/10.1007/s00146-025-02207-y

Kiiski, H. S. M., Cullen, B., Clavin, S. L., & Newell, F. N. (2016). Perceptual and Social Attributes Underlining Age-Related Preferences for Faces. *Frontiers in Human Neuroscience*, *10*. https://www.frontiersin.org/articles/10.3389/fnhum.2016.00437

Lawrence, M., Cimermanis, K. N. E., & Collin, C. A. (2025). Not all AI-generated faces are created equal: impacts of model gender, race, and emotional expression on classification accuracy. *AI & SOC*. https://doi.org/10.1007/s00146-025-02670-7

Lawrence, M. & Collin, C. A. (2025). Recognition of emotional expressions in real vs. AI-generated face images: Effects of race, gender, and expression. [Manuscript submitted].

Lee, K. (2013). Why Asian female stereotypes matter to all: Beyond black and white, east and west. *Critical Philosophy of Race, 1*(1), 86-103.

Liefooghe, B., Oliveira, M., Leisten, L. M., Hoogers, E., Aarts, H., & Hortensius, R. (2023). Are Natural Faces Merely Labelled as Artificial Trusted Less? *Collabra: Psychology*, *9*(1), 73066.

Locke, L. G., & Hodgdon, G. (2024). Gender bias in visual generative artificial intelligence systems and the socialization of AI. *AI & SOCIETY*, 1-8. https://doi.org/10.1007/s00146-024-02129-1

Markey, P. M., & Markey, C. N. (2009). A Brief Assessment of the Interpersonal Circumplex: The IPIP-IPC. *Assessment*, *16*(4), 352–361. https://doi.org/10.1177/1073191109340382

Melson-Silimon, A., Spivey, B. N., & Skinner-Dorkenoo, A. L. (2024). The construction of racial stereotypes and how they serve as racial propaganda. *Social and Personality Psychology Compass, 18*(1), e12862. DOI: 10.1111/spc3.12862

Morgan, M., & Bennett, D. (2006). Getting off of Black women's backs: Love her or leave her alone. *Du Bois Review: Social Science Research on Race, 3*(2), 485-502. https://doi.org/10.1017/S1742058X06060334

Mubashir, M. (2024). The gendered dress of DALL-E 2: Exploring profession-based images in the Indian context. *MedieKultur: Journal of media and communication research*, *40*(76), 100-119. https://doi.org/10.7146/mk.v40i76.143565

Nguyen, C. F. (2016). Asian American women faculty: Stereotypes and triumphs.

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

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Park, Y. S. (2024). White default: Examining racialized biases behind AI-generated images. *Art Education*, *77*(4), 36-45. https://doi.org/10.1080/00043125.2024.2330340

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343. https://doi.org/10.1098/rsos.140343

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

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

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

Walley-Jean, J. C. (2009). Debunking the myth of the "angry Black woman": An exploration of anger in young African American women. *Black Women, Gender & Families*, *3*(2), 68-86. https://www.jstor.org/stable/10.5406/blacwomegendfami.3.2.0068

Wiggins, J. S., Trapnell, P., & Phillips, N. (1988). Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research*, *23*(4), 517–530. https://doi.org/10.1207/s15327906mbr2304_8

Wu, Y., Nakashima, Y., & Garcia, N. (2025). Revealing gender bias from prompt to image in stable diffusion. *Journal of Imaging*, *11*(2), 35. DOI: 10.3390/jimaging11020035

Zebrowitz, L. A. (2017). First Impressions From Faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996

Zhang, D., Lin, H., & Perrett, D. I. (2020). Apparent Emotional Expression Explains the Effects of Head Posture on Perceived Trustworthiness and Dominance, but a Measure of Facial Width Does Not. *Perception*, *49*(4), 422–438. https://doi.org/10.1177/0301006620909286

## 4.9. Supplementary Materials

### Table S1

*Demographic Characteristics of the Total Sample*

| | Mean | SD |
|---|---|---|
| Age (Years) | 19.10 | 4.40 |
| Education (Years) | 1.41 | .92 |
| Time spent in Canada (Years) | 5.93 | 4.63 |
| | N | Percentage (%) |
| Race | | |
| Arab | 3 | 3.7 |
| Arab/African | 3 | 3.7 |
| Asian-East | 7 | 8.6 |
| Asian-South | 8 | 9.9 |
| Asian-Southeast | 4 | 4.9 |
| Black-African | 15 | 18.5 |
| Black-Caribbean | 3 | 3.7 |
| Black-North American | 3 | 3.7 |
| First Nation | 0 | 0 |
| Indian-Caribbean | 0 | 0 |
| Indigenous/Aboriginal | 0 | 0 |
| Inuit | 0 | 0 |
| Latin American | 2 | 2.5 |
| Metis | 0 | 0 |
| Middle Eastern | 7 | 8.6 |
| Mixed Heritage | 3 | 3.7 |
| Pacific Islander | 0 | 0 |
| White-North American | 33 | 40.7 |
| White-European | 3 | 3.7 |
| You don’t have an option | 1 | 1.2 |

| | | |
|---|---|---|
| that applies to me | | |
| Choose not to respond | 0 | 0 |
| Gender | | |
| Agender | 2 | 2.5 |
| Cisgender Female | 61 | 75.3 |
| Cisgender Male | 18 | 22.2 |
| Genderfluid | 1 | 1.2 |
| Genderqueer | 1 | 1.2 |
| Non-binary | 1 | 1.2 |
| Questioning | 1 | 1.2 |
| Transgender Female | 0 | 0 |
| Transgender Male | 0 | 0 |
| Two Spirit | 0 | 0 |
| No Option Applies | 0 | 0 |
| Choose not to respond | 0 | 0 |

*Note.* Mean and Standard Deviation (SD) for age (n = 80), years of education (n = 81), and time spent in Canada (n = 30) for participants. For time spent in Canada, only those who reported living somewhere else were included. Number of participants (N) and percentage of race and gender of participants (N = 81). The total percentage for race and gender may not equal 100% as participants were instructed to select all that apply, so some participants selected more than one option, or due to missing data. Although Mixed Heritage was an option, not all participants who selected multiple options selected this, and not all participants who selected Mixed Heritage specified further.

**Table S2**

*Omnibus ANOVA Significant Interactions and Main Effects for Emotional Valence*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | F(1, 76) = 57.30 | < .001 | .43 |
| Gender | F(1, 76) = 55.87 | < .001 | .42 |
| Race | F(2, 152) = 44.57 | < .001 | .37 |
| Emotion | F(1, 76) = 2083.64 | < .001 | .97 |
| Image Type*Gender | F(1, 76) = 9.69 | .003 | .11 |
| Image Type*Race | F(2, 152) = 11.32 | < .001 | .13 |
| Image Type*Emotion | F(1, 76) = 4.85 | .03 | .06 |
| Gender*Race | F(1.75, 133.05) = 20.67 | < .001 | .21 |
| Race*Emotion | F(2, 152) = 4.06 | .02 | .05 |
| Image Type*Gender*Race | F(2, 152) = 26.62 | < .001 | .26 |
| Image Type*Gender*Emotion | F(1, 76) = 139.06 | < .001 | .65 |
| Gender*Race*Emotion | F(2, 152) = 37.52 | < .001 | .33 |
| Image Type*Gender*Race*Emotion | F(2, 152) = 9.49 | < .001 | .11 |

*Note.* N = 79

**Table S3**

*Omnibus ANOVA Significant Interactions and Main Effects for Arousal*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | F(1, 78) = 154.52 | < .001 | .67 |
| Gender | F(1, 78) = 30.83 | < .001 | .28 |
| Emotion | F(1, 78) = 210.97 | < .001 | .73 |
| Image Type*Gender | F(1, 78) = 44.53 | < .001 | .36 |
| Image Type*Race | F(2, 156) = 16.82 | < .001 | .18 |
| Image Type*Emotion | F(1, 78) = 47.63 | < .001 | .38 |
| Gender*Emotion | F(1, 78) = 6.76 | .01 | .08 |
| Race*Emotion | F(2, 156) = 7.93 | < .001 | .09 |
| Image Type*Gender*Race | F(2, 156) = 17.02 | < .001 | .18 |
| Image Type*Gender*Race*Emotion | F(2, 156) = 14.58 | < .001 | .16 |

*Note.* N = 81

**Table S4**

*Omnibus ANOVA Significant Interactions and Main Effects for Warmth*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | F(1, 78) = 23.80 | < .001 | .23 |
| Gender | F(1, 78) = 29.70 | < .001 | .28 |
| Race | F(2, 156) = 3.10 | .05 | .04 |
| Emotion | F(1, 78) = 659.40 | < .001 | .89 |
| Image Type*Gender | F(1, 78) = 7.32 | .008 | .09 |
| Image Type*Race | F(2, 156) = 16.16 | < .001 | .17 |
| Gender*Race | F(2, 156) = 28.91 | < .001 | .27 |
| Image Type*Emotion | F(1, 78) = 52.78 | < .001 | .40 |
| Race*Emotion | F(2, 156) = 9.84 | < .001 | .11 |
| Image Type*Gender*Race | F(2, 156) = 16.53 | < .001 | .18 |
| Image Type*Gender*Emotion | F(1, 78) = 57.28 | < .001 | .42 |
| Image Type*Race*Emotion | F(2, 156) = 15.63 | < .001 | .17 |
| Gender*Race*Emotion | F(2, 156) = 14.13 | < .001 | .15 |
| Image Type*Gender*Race*Emotion | F(2, 156) = 3.42 | .04 | .04 |

*Note.* N = 81

**Table S5**

*Omnibus ANOVA Significant Interactions and Main Effects for Dominance*

| IV | *F*-value | *p*-value | Effect Size ($\eta_p^2$) |
|---|---|---|---|
| Image Type | F(1, 77) = 60.07 | < .001 | .44 |
| Gender | F(1, 77) = 7.84 | .006 | .09 |
| Race | F(2, 154) = 23. 58 | < .001 | .23 |
| Emotion | F(1, 77) = 137.32 | < .001 | .64 |
| Image Type*Gender | F(1, 77) = 22.82 | < .001 | .23 |
| Image Type*Race | F(2, 154) = 45.84 | < .001 | .37 |
| Image Type*Emotion | F(1, 77) = 126.49 | < .001 | .62 |
| Gender*Race | F(2, 154) = 49.37 | < .001 | .39 |
| Gender*Emotion | F(1, 77) = 11.06 | .001 | .13 |
| Race*Emotion | F(2, 154) = 9.92 | < .001 | .1 |
| Image Type*Gender*Race | F(2, 154) = 15.65 | < .001 | .17 |
| Image Type*Gender*Emotion | F(1, 77) = 17.62 | < .001 | .19 |
| Image Type*Race*Emotion | F(2, 154) = 31.32 | < .001 | .29 |
| Gender*Race*Emotion | F(1.78, 137.14) = 37.78 | < .001 | .33 |

*Note.* N = 80

# 5. GENERAL DISCUSSION

AI image generation is a new and rapidly evolving technology. What was started by hobbyists creating clearly fake images and videos (Whittaker et al., 2020), is becoming a torrent of sometimes imperceptible artificial images (Bray et al., 2023; Josephs et al., 2023; Moshel et al., 2022; Nightingale & Farid, 2022; Rossi et al., 2023; Tucciarelli et al., 2022). With documented cases of criminal misuse of AI image generation (FBI, 2023) and the potential for a wide range of other misuses (Whittaker et al., 2020), being able to understand how and when we are able to distinguish between real and AI-generated media is paramount. Moreover, because AI-generated images also have positive uses (Eqbal, 2023; Whittaker et al., 2020), it is important to understand to what degree they effectively convey social and emotional information. We make inferences about others based on their faces and emotional expressions (Olivola et al., 2014; Todorov et al., 2015; Van Kleef, 2009). We also have affective reactions towards others based on the emotions we have decoded from their facial expressions. These both guide our behaviours towards them. As AI generation can be used in positive situations, for example, eliminating language barriers (Eqbal, 2023; Whittaker et al., 2020), understanding its ability to accurately recreate emotion is necessary. However, whether the inferences we make are comparable for real and AI-generated faces is understudied. Importantly, the studies described above used real face photographs of individuals posing emotional expressions, rather than natural expressions. While the stimuli were previously validated, this distinction is important as the AI-generator, particularly the refined models, would be drawing, at least partially, on posed facial expressions, rather than genuine expressions of emotion.

Another understudied aspect of AI image generation is how it interacts with gender and race. These issues have already been examined extensively in real faces, but little work has been

done examining how AI-generated and photographic faces compare. Work with photographs shows that women may have an advantage over men in emotional expression recognition (Kret & De Gelder, 2012; McClure, 2000; Montagne et al., 2005). Beyond differences in emotion recognition between men and women, how emotion interacts with model gender has been examined extensively as well. For example, displays of certain emotions may be perceived as more feminine or masculine (Hess et al., 2009). Another study found that men's faces must display more happiness than women's faces to be perceived as neutral, indicating that men's faces may be perceived as more negative than women's faces even when the intensity of the emotion is equal (Harris et al., 2016). Women's faces also tend to be rated as more approachable than men's faces (Campbell et al., 2010).

The race of the participant and model may interact when one is trying to recognize emotions (Beaupré & Hess, 2005; Campbell et al., 2010; Jiang et al., 2023; Young et al., 2011). While the work examining whether there is a racial ingroup advantage for emotion recognition is mixed, the work does consistently find that race impacts emotion recognition. For example, Beaupré and Hess (2005) did not find an overall ingroup advantage, but they did find that African decoders identified expressions of fear more accurately when the emotion was displayed by an ingroup member rather than an outgroup member. Moreover, they found differences in decoder race and accuracy in identifying certain emotions. Campbell et al. (2010) examined differences in approachability of faces for men and women, and White and non-White faces. When they controlled for imblanaces in the stimuli, they found that White faces were rated as more approachable than non-White faces. They also found that race and gender of the face interacted, whereby non-White men's faces were rated as more approachable than White men's faces, with no difference for women's faces. Broadly speaking, emotional and social judgements

can be influenced by a variety of observer and model characteristics like gender, race, and age, and these have been extensively studied in the real-face literature (Beaupré & Hess, 2005; Campbell et al., 2010; Cortes et al., 2019; Harris et al., 2016; Hess et al., 2009; Jiang et al., 2023; Kiiski et al., 2016; Olivola et al., 2014; Young et al., 2011; Zebrowitz, 2017; Zhang et al., 2020) but not in the literature for AI-generated media.

The three studies outlined above helped to expand this area of research by (i) examining whether people can accurately classify real vs. AI-generated images of emotionally expressive faces, (ii) examining whether emotion recognition is as accurate for AI-generated faces when compared to real faces, (iii) examining the saliency of emotions in AI-generated compared to real faces, (iv) examining the intensity of emotions of AI-generated compared to real faces, and (v) examining the differences in social attributions assigned to AI-generated compared to real faces. All of these questions were examined across gender and race of the models, with some exploratory analyses based on participant characteristics.

In Study 1, we explored whether using a refined AI-generator would reduce or eliminate some of the gender and racial biases other studies have found (Nightingale & Farid, 2022). We also included a variety of emotional expressions, an area which has largely been neglected. Despite using a refined model that was weighted towards inclusion of POC, women and multiple emotional expressions, we still found differences based on these variables. Overall, AI-generated faces were quite detectable. Specifically, Asian and Black AI-generated faces were consistently more easily detected than White faces. However, this finding interacted with gender and emotional expression such that Asian men were more detectable across all emotions, but Black men were more detectable for anger expressions only. Moreover, Black women were most detectable in the happy and sad expressions. These findings support the idea that, despite our

best efforts being made to generate realistic images of POC, the training sets are so strongly weighted towards White individuals that these biases come out in the stimuli anyway. Based on our findings, we also see certain stereotypes being perpetuated. For example, AI-generated images of Black women are more detectable in the happy and fear emotional expression conditions, but not the anger and sad conditions; this indicates that the latter expressions for Black women are more realistic. This could be due to the stereotype of the "angry Black woman", which is a pervasive stereotype classifying Black women as angry, aggressive, loud, and hostile, despite no empirical evidence to support this notion (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009).

Surprisingly, we found that AI-generated images of men's faces were more detectable than those of women's faces (again, this interacted with race and emotional expression). This is contrary to what we expected because many other studies show evidence supporting an over-representation of men in AI training sets (Bhardwaj et al., 2025; Chauhan et al., 2024; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Locke & Hodgdon, 2025; Nightingale & Farid, 2022; Wu et al., 2025).

In study 2, we examined how the effects of race, gender, and emotional expression of faces interact when comparing AI-generated and real photographs regarding the impact of their emotional expressions. Overall, we saw a trend of higher accuracy, saliency, and intensity ratings for White faces compared to Asian and Black faces, indicating, again, an over-representation of White faces in the training sets. In addition, we found that happy faces by far produced the best accuracy, saliency, and intensity ratings, even though the AI-generated faces were still rated lower on all these measures than their real face counterparts. Also worth noting is that this same pattern of lower emotional amplitude in AI-generated faces appeared across all race and gender

combinations, though the effect was stronger for men. As in study 1 (Lawrence, Cimermanis, & Collin, 2025), and contrary to other studies (Bhardwaj et al., 2025; Chauhan et al., 2024; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Locke & Hodgdon, 2025; Nightingale & Farid, 2022; Wu et al., 2025), this would seem to indicate an *under*-representation of men in training sets rather than *over*-representation and therefore should be studied further.

When examining how gender, race, and emotional expression interacted to produce different effects for accuracy of emotion detection, saliency, and intensity, we again found certain stereotypes being reflected. Again, we saw evidence of the AI-generator producing the stereotype of the "angry Black woman" (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009). For all three dependent variables, Black women expressing anger were more accurately identified, and the emotion was more salient and intense compared to the real face counterpart, whereas for all other emotional expressions, they were less or equally accurately identified, salient, and intense as their real face counterparts.

In study 3, results indicated several different stereotypes being reflected by the AI-generator. Firstly, as in studies 1 and 2, we saw reproduction of the "angry Black woman" (Ashley, 2014; Lee, 2013; Melson-Silimon et al., 2024; Morgan & Bennett, 2006; Walley-Jean, 2009). This stereotype emerged in the ratings of negative valence, coldness, and dominance for the AI-generated faces of Black women expressing anger being much stronger than those of their real face counterparts, or than any other women's faces in the study. These faces were more often described using terms expressing negativity, aggression, and unapproachability. Secondly, we saw evidence of Black faces (regardless of gender) being perceived as more hostile than other faces (Melson-Silimon et al., 2024). This was evidenced by AI-generated faces of Black people

being rated as equally dominant to their real face counterparts, whereas for Asian and White AI-generated faces, they were lower on dominance compared to their real face counterparts.

Another stereotype that emerged was the "passive/submissive Asian woman" (Azhar et al., 2021; Lee, 2013; Nguyen, 2016), whereby Asian women's faces were consistently closer to the neutral point (or higher in the case of warmth) compared to their real face counterparts and compared to the AI-generated faces of Black and White women. We also see a related stereotype of "weak Asian men" (Azhar et al., 2021) being perpetuated where, when expressing anger, Asian men's faces were rated warmer than both their real counterparts and AI-generated faces of Black and White men. Moreover, for both anger and happiness, we saw reduced levels of dominance. Finally, we saw both of these stereotypes evidenced when collapsing across gender, and seeing that the AI-generated images of Asian faces were more submissive (i.e., less dominant) than their real face counterparts, and AI-generated faces of Black and White individuals, despite dominance levels being similar for the three races across real face images.

## 6. GENERAL CONCLUSION

Taken all together, the three studies outlined above confirm and add new evidence to the notion that AI-generators, particularly those implementing Stable Diffusion, are prone to replicating and enhancing societally held stereotypes and biases (AlDahoul et al., 2025; Assis & Moura, 2025; Bianchi et al., 2023; Chauhan et al., 2024; Gawali et al., 2024; Ghosh & Caliskan, 2023; Gorska & Jemielniak, 2023; Jääskeläinen et al., 2025; Lawrence, Carriere, & Collin, 2025; Lawrence, Cimermanis, & Collin, 2025; Lawrence & Collin, 2025; Mubashir, 2024; Park, 2024). This occurs due to the biased training sets used to create generators. This has many negative implications for real world uses of AI-generation, and related software such as facial recognition. As AI becomes more advanced, prevalent, and difficult to detect, it will continue contributing to

the negative view of different groups. As the prevalence of AI-generated media continues to increase, this work underscores the importance of training AI-generators using balanced image sets, for example, the Fairface dataset (Kärkkäinen & Joo, 2021), which has been balanced across gender (men and women), race (White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino), and age. AI-generated media will be biased based on the data they are trained on. If this media is going to be present in our day-to-day lives, it is important that we take steps to reduce the negative stereotypes being replicated and perpetuated.

# 7. REFERENCES

AlDahoul N, Rahwan T, Zaki Y (2025) Ai-generated faces influence gender stereotypes and racial homogenization. Sci Rep 15(1):14449. https://doi.org/10.1038/s41598-025-99623-3

American Psychological Association, APA Task Force on Race and Ethnicity Guidelines in Psychology. (2019). Race and Ethnicity Guidelines in Psychology: Promoting Responsiveness and Equity. Retrieved from http://www.apa.org/about/policy/race-and-ethnicity-in-psychology.pdf

Ames, D. R., & Bianchi, E. C. (2008). The Agreeableness Asymmetry in First Impressions: Perceivers' Impulse to (Mis)judge Agreeableness and How It Is Moderated by Power. *Personality and Social Psychology Bulletin*, *34*(12), 1719–1736. https://doi.org/10.1177/0146167208323932

Anderson, A. K., Yamaguchi, Y., Grabski, W., & Lacka, D. (2006). Emotional memories are not all created equal: Evidence for selective memory enhancement. *Learning & Memory*, *13*(6), 711–718. https://doi.org/10.1101/lm.388906

Ashley, W. (2014). The Angry Black Woman: The Impact of Pejorative Stereotypes on Psychotherapy with Black Women. *Social Work in Public Health*, *29*(1), 27–34. https://doi.org/10.1080/19371918.2011.619449

Assis, J. de, & Moura, M. A. (2025). Algorithmic Semiosis and Racial Bias: A Study of Images Created by Generative AI. *Encontros Bibli*, *30*, e103495. https://doi.org/10.5007/1518-2924.2025.e103495

Azhar, S., Alvarez, A. R. G., Farina, A. S. J., & Klumpner, S. (2021). “You’re So Exotic Looking”: An Intersectional Analysis of Asian American and Pacific Islander Stereotypes. *Affilia*, *36*(3), 282–301. https://doi.org/10.1177/08861099211001460

Baudouin, J.-Y., Gilibert, D., Sansone, S., & Tiberghien, G. (2000). When the smile is a cue to familiarity. *Memory*, *8*(5), 285–292. https://doi.org/10.1080/09658210050117717

Beaupré, M. G., & Hess, U. (2005). Cross-Cultural Emotion Recognition among Canadian Ethnic Groups. *Journal of Cross-Cultural Psychology*, *36*(3), 355–370. https://doi.org/10.1177/0022022104273656

Bhardwaj, N., Bhardwaj, A., & Garg, L. (2025). Controlling Bias in Generative AI: Techniques for Fair and Equitable Data Generation in Socially Sensitive Applications. *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 1–8. https://doi.org/10.1109/ICICACS65178.2025

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. https://doi.org/10.1145/3593013.3594095

Boutet, I., Guay, J., Chamberland, J., Cousineau, D., & Collin, C. (2023). Emojis that work! Incorporating visual cues from facial expressions in emojis can reduce ambiguous interpretations. *Computers in Human Behavior Reports*, *9*, 100251. https://doi.org/10.1016/j.chbr.2022.100251

Bradley, B. P., Mogg, K., Millar, N., Bonham-Carter, C., Fergusson, E., Jenkins, J., & Parr, M. (1997). Attentional biases for emotional faces. *Cognition & Emotion*, *11*(1), 25–42.

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake'images of human faces. *Journal of Cybersecurity*, *9*(1), tyad011. https://doi.org/10.1093/cybsec/tyad011

Campbell, D. W., Neuert, T., Friesen, K. B., & McKeen, N. A. (2010). Assessing Social Approachability: Individual Differences, In-Group Biases, and Experimental Control. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *42*(4), 254–263. https://doi.org/10.1037/a0020229

Carstensen, L. L., & DeLiema, M. (2018). The positivity effect: A negativity bias in youth fades with age. *Current Opinion in Behavioral Sciences*, *19*, 7–12. https://doi.org/10.1016/j.cobeha.2017.07.009

Chandaliya, P. K., & Nain, N. (2022). ChildGAN: Face aging and rejuvenation to find missing children. *Pattern Recognition*, *129*, 108761. https://doi.org/10.1016/j.patcog.2022.108761

Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and emotional memory: The forgettable nature of negative images for older adults. *Journal of Experimental Psychology: General*, *132*(2), 310–324. https://doi.org/10.1037/0096-3445.132.2.310

Chauhan, A., Anand, T., Jauhari, T., Shah, A., Singh, R., Rajaram, A., & Vanga, R. (2024). Identifying race and gender bias in stable diffusion ai image generation. *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, 1–6. https://doi.org/10.1109/ICAIC60265.2024.10433840

Chen, W., Liu, C. H., Li, H., Tong, K., Ren, N., & Fu, X. (2015). Facial expression at retrieval affects recognition of facial identity. *Frontiers in Psychology*, *6*, 780. https://doi.org/10.3389/fpsyg.2015.00780

Collin, C. A., Chamberland, J., LeBlanc, M., Ranger, A., & Boutet, I. (2022). Effects of emotional expression on face recognition may be accounted for by image similarity. Social Cognition, 40(3), 282-301. https://doi.org/10.1521/soco.2022.40.3.282

Conley, M. I., Dellarco, D. V., Rubien-Thomas, E., Cohen, A. O., Cervera, A., Tottenham, N., & Casey, B. (2018). The racially diverse affective expression (RADIATE) face stimulus set. *Psychiatry Research*, *270*, 1059–1067. https://doi.org/10.1016/j.psychres.2018.04.066

Cortes, D. S., Link to external site,  this link will open in a new window, Laukka, P., Link to external site,  this link will open in a new window, Ebner, N. C., & Fischer, H. (2019). Age-related differences in evaluation of social attributes from computer-generated faces of varying intensity. *Psychology and Aging*, *34*(5), 686–697. https://doi.org/10.1037/pag0000364

Dehouche, N. (2021). Implicit stereotypes in pre-trained classifiers. *IEEE Access*, *9*, 167936–167947. https://doi.org/10.1109/ACCESS.2021.3136898

Dhariwal P, Nichol A (2021) Diffusion models beat GANs on image synthesis. Adv Neural Inf Process Syst 34:8780–8794

Duffy, S., August, A., & Wisniewski, K. (2024). Discriminating real from AI-generated faces: Effects of emotion, gender, and age. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/49v4z44k

Eiserbeck, A., Maier, M., Baum, J., & Abdel Rahman, R. (2023). Deepfake smiles matter less—The psychological and neural impact of presumed AI-generated faces. *Scientific Reports*, *13*(1), 16111. https://doi.org/10.1038/s41598-023-42802-x

Ekman, P. (1989). The argument and evidence about universals in facial expressions. *Handbook of Social Psychophysiology*, 143–164.

Ekman, P. (1992). *Are there basic emotions? 99*(3), 550–553.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*. https://psycnet.apa.org/doiLanding?doi=10.1037/t27734-000.

Eqbal, A. (2023, April 12). *What if you could use AI to fix the lips in dubbed movies? This Hollywood director is doing it | CBC Arts*. CBC. https://www.cbc.ca/arts/commotion/what-if-you-could-use-ai-to-fix-the-lips-in-dubbed-movies-this-hollywood-director-is-doing-it-1.6808419

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

FBI (Federal Bureau of Investigation). (2023, June 5). *Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. https://www.ic3.gov/Media/Y2023/PSA230605

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967–984. https://doi.org/10.1037/0022-3514.74.4.967

Fraser, K. C., Kiritchenko, S., & Nejadgholi, I. (2023). *A Friendly Face: Do Text-to-Image Systems Rely on Stereotypes when the Input is Under-Specified?* (arXiv:2302.07159). arXiv. http://arxiv.org/abs/2302.07159

Gawali, A., Tarle, A., Can Inac, C., & Kulkarni, S. S. (2024). *A Framework for Detecting Stereotypes, Prejudices and Discrimination in AI-Generated Imagery*. https://www.researchgate.net/profile/Can-Inac/publication/388105344_A_Framework_for_Detecting_Stereotypes_Prejudices_and_Discrimination_in_AI-Generated_Imagery/links/678a22c895e02f182e977a29/A-Framework-for-Detecting-Stereotypes-Prejudices-and-Discrimination-in-AI-Generated-Imagery.pdf

Ghosh, S., & Caliskan, A. (2023). 'Person' == Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6971–6985. https://doi.org/10.18653/v1/2023.findings-emnlp.465

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Gorska, A. M., & Jemielniak, D. (2023). The invisible women: Uncovering gender bias in AI-generated images of professionals. *Feminist Media Studies*, *23*(8), 4370–4375. https://doi.org/10.1080/14680777.2023.2263659

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Goulet, M.-A., & Cousineau, D. (2019). The Power of Replicated Measures to Increase Statistical Power. *Advances in Methods and Practices in Psychological Science*, *2*(3), 199–213. https://doi.org/10.1177/2515245919849434

Grady, C. L., Hongwanishkul, D., Keightley, M., Lee, W., & Hasher, L. (2007). The effect of age on memory for emotional faces. *Neuropsychology*, *21*(3), 371. https://doi.org/10.1037/0894-4105.21.3.371

Gragnaniello, D., Marra, F., & Verdoliva, L. (2022). Detection of AI-Generated Synthetic Faces. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks* (pp. 191–212). Springer International Publishing Cham.

Gurtman, M. B., & Pincus, A. L. (2000). Interpersonal Adjective Scales: Confirmation of Circumplex Structure from Multiple Perspectives. *Personality and Social Psychology Bulletin*, *26*(3), 374–384. https://doi.org/10.1177/0146167200265009

Harris, D. A., Hayes-Skelton, S. A., & Ciaramitaro, V. M. (2016). What's in a face? How face gender and current affect influence perceived emotion. *Frontiers in psychology, 7*, 1468. https://doi.org/10.3389/fpsyg.2016.01468

Hess, U., Adams, R. B., Grammer, K., & Kleck, R. E. (2009). Face gender and emotion expression: Are angry women more like men?. *Journal of Vision, 9*(12), 19-19. https://doi.org/10.1167/9.12.19

*Home | Leonardo.Ai*. (n.d.). Leonardo.Ai. https://app.leonardo.ai/

Jääskeläinen, P., Sharma, N. K., Pallett, H., & Åsberg, C. (2025). Intersectional analysis of visual generative AI: The case of stable diffusion. *AI & SOCIETY*. https://doi.org/10.1007/s00146-025-02207-y

Jackson, M. C., Wolf, C., Johnston, S. J., Raymond, J. E., & Linden, D. E. (2008). Neural correlates of enhanced visual short-term memory for angry faces: An FMRI study. *PloS One*, *3*(10), e3536. https://doi.org/10.1371/journal.pone.0003536

Jiang, Z., Recio, G., Li, W., Zhu, P., He, J., & Sommer, W. (2023). The other-race effect in facial expression processing: Behavioral and ERP evidence from a balanced cross-cultural study in women. *International Journal of Psychophysiology*, *183*, 53–60. https://doi.org/10.1016/j.ijpsycho.2022.11.009

Josephs, E., Fosco, C., & Oliva, A. (2023). Artifact magnification on deepfake videos increases human detection and subjective confidence. *Journal of Vision (Charlottesville, Va.)*, *23*(9), 5327. https://doi.org/10.1167/jov.23.9.5327

Kärkkäinen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. Proceedings/IEEE Workshop on Applications of Computer Vision, 1547–1557. https://doi.org/10.1109/WACV48630.2021.00159

Kiiski, H. S. M., Cullen, B., Clavin, S. L., & Newell, F. N. (2016). Perceptual and Social Attributes Underlining Age-Related Preferences for Faces. *Frontiers in Human Neuroscience*, *10*. https://www.frontiersin.org/articles/10.3389/fnhum.2016.00437

Kret, M. E., & De Gelder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia*, *50*(7), 1211–1221. https://doi.org/10.1016/j.neuropsychologia.2011.12.022

Lago, F., Pasquini, C., Bohme, R., Dumont, H., Goffaux, V., & Boato, G. (2022). More Real Than Real: A Study on Human Visual Perception of Synthetic Faces [Applications Corner]. *IEEE Signal Processing Magazine*, *39*(1), 109–116. https://doi.org/10.1109/MSP.2021.3120982

Lawrence, M., Carriere, E., & Collin, C. A. (2025). Social Attributions Given to Happy and Angry AI-Generated and Real Faces: Is AI Reproducing Race and Gender Biases? [Manuscript Submitted].

Lawrence, M., Cimermanis, K. N. E., & Collin, C. A. (2025). Not all AI-generated faces are created equal: impacts of model gender, race, and emotional expression on classification accuracy. *AI & SOC*, 1-12. https://doi.org/10.1007/s00146-025-02670-7

Lawrence, M. & Collin, C. A. (2025). Recognition of emotional expressions in real vs. AI-generated face images: Effects of race, gender, and expression. [Manuscript submitted].

Lee, K. (2013). Why Asian female stereotypes matter to all: Beyond black and white, east and west. *Critical Philosophy of Race*, *1*(1), 86–103. https://doi.org/10.5325/critphilrace.1.1.0086

Liefooghe, B., Oliveira, M., Leisten, L. M., Hoogers, E., Aarts, H., & Hortensius, R. (2023). Are Natural Faces Merely Labelled as Artificial Trusted Less? *Collabra: Psychology*, *9*(1), 73066. https://doi.org/10.1525/collabra.73066

Liu, C. H., Chen, W., & Ward, J. (2014). Remembering faces with emotional expressions. *Frontiers in Psychology*, *5*, 1439. https://doi.org/10.3389/fpsyg.2014.01439

Locke, L. G., & Hodgdon, G. (2025). Gender bias in visual generative artificial intelligence systems and the socialization of AI. *AI & SOCIETY*, *40*(4), 2229–2236. https://doi.org/10.1007/s00146-024-02129-1

Markey, P. M., & Markey, C. N. (2009). A Brief Assessment of the Interpersonal Circumplex: The IPIP-IPC. *Assessment*, *16*(4), 352–361. https://doi.org/10.1177/1073191109340382

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin*, *126*(3), 424–453. https://doi.org/10.1037/0033-2909.126.3.424

Melson-Silimon, A., Spivey, B. N., & Skinner-Dorkenoo, A. L. (2024). The construction of racial stereotypes and how they serve as racial propaganda. *Social and Personality Psychology Compass*, *18*(1), e12862. https://doi.org/10.1111/spc3.12862

Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A. M., Krumhuber, E. G., & Dawel, A. (2023). AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, *34*(12), 1390–1403. https://doi.org/10.1177/09567976231207095

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. https://doi.org/10.1007/s10339-005-0050-6

Morgan, M., & Bennett, D. (2006). Getting off of Black women's backs: Love her or leave her alone. *Du Bois Review: Social Science Research on Race*, *3*(2), 485–502. https://doi.org/10.1017/S1742058X06060334

Moshel, M. L., Robinson, A. K., Carlson, T. A., & Grootswagers, T. (2022). Are you for real? Decoding realistic AI-generated faces from neural activity. *Vision Research*. https://doi.org/10.1016/j.visres.2022.108079

Mubashir, M. (2024). The gendered dress of DALL-E 2: Exploring profession-based images in the Indian context. *MedieKultur: Journal of Media and Communication Research*, *40*(76), 100–119. https://doi.org/10.7146/mk.v40i76.143565

Nguyen, C. F. (2016). *Asian American women faculty: Stereotypes and triumphs*. https://repository.usfca.edu/listening_to_the_voices/12/

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

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. https://doi.org/10.1016/j.tics.2014.09.007

Patil, S., Cuenca, P., Lambert, N., & von Platen, P. (2022, August 22). *Stable Diffusion with 🧨 Diffusers*. Hugging Face. https://huggingface.co/blog/stable_diffusion

Park, Y. S. (2024). White Default: Examining Racialized Biases Behind AI-Generated Images. *Art Education*, *77*(4), 36–45. https://doi.org/10.1080/00043125.2024.2330340

Quigley, J. (2023). *The Call from Inside the House: The Final Girl Trope as a Reflection of Cultural Anxiety of Race and Gender* (Master's thesis, Idaho State University).

Rossi, S., Kwon, Y., Auglend, O. H., Mukkamala, R. R., Rossi, M., & Thatcher, J. (2023). Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?

*Proceedings of the 56th Hawaii International Conference on System Sciences,* 134–143. https://hdl.handle.net/10125/102645. Accessed 16 Sept 2025.

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, *2*(6), 140343. https://doi.org/10.1098/rsos.140343

Shen, B., RichardWebster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021). A Study of the Human Perception of Synthetic Faces. *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8. https://doi.org/10.1109/FG52635.2021.9667066

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

Tay, P. K., & Yang, H. (2017). Angry faces are more resistant to forgetting than are happy faces: Directed forgetting effects on the identity of emotional faces. *Journal of Cognitive Psychology*, *29*(7), 855–865. https://doi.org/10.1080/20445911.2017.1323907

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, *66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B., & Nelson, C. (2009). The NimStim set of facial expressions:

Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

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

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science : A Journal of the American Psychological Society, 18*(3), 184–188. https://doi.org/10.1111/j.1467-8721.2009.01633.x

Van Selst, M., & Jolicoeur, P. (1994). A Solution to the Effect of Sample Size on Outlier Elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. https://doi.org/10.1080/14640749408401131

Walley-Jean, J. C. (2009). Debunking the myth of the "angry Black woman": An exploration of anger in young African American women. *Black Women, Gender & Families*, *3*(2), 68–86. https://doi.org/10.1353/bwg.0.0011

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F.-Y. (2017). Generative adversarial networks: Introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, *4*(4), 588–598. IEEE/CAA Journal of Automatica Sinica. https://doi.org/10.1109/JAS.2017.7510583

Wang, X., Guo, H., Hu, S., Chang, M.-C., & Lyu, S. (2022). Gan-generated faces detection: A survey and new perspectives. *arXiv Preprint arXiv:2202.07145*.

Wang, Y., Zhu, Z., Chen, B., & Fang, F. (2019). Perceptual learning and recognition confusion reveal the underlying relationships among the six basic emotions. *Cognition and Emotion*, *33*(4), 754–767. https://doi.org/10.1080/02699931.2018.1491831

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": The mad world of AI-generated media. *IT Professional*, *22*(5), 90–99. https://doi.org/10.1109/MITP.2020.2985492

Wiggins, J. S., Trapnell, P., & Phillips, N. (1988). Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research*, *23*(4), 517–530. https://doi.org/10.1207/s15327906mbr2304_8

Wu, Y., Nakashima, Y., & Garcia, N. (2025). Revealing gender bias from prompt to image in stable diffusion. *Journal of Imaging*, *11*(2), 35. https://doi.org/10.3390/jimaging11020035

Yan, X., Andrews, T. J., Jenkins, R., & Young, A. W. (2016). Cross-cultural differences and similarities underlying other-race effects for facial identity and expression. *Quarterly Journal of Experimental Psychology*, *69*(7), 1247–1254. https://doi.org/10.1080/17470218.2016.1146312

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2011). Perception and motivation in face recognition: A critical review of theories of the Cross-Race Effect. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *16*(2), 116–142. https://doi.org/10.1177/1088868311418987

Zebrowitz, L. A. (2017). First Impressions From Faces. *Current Directions in Psychological Science*, *26*(3), 237–242. https://doi.org/10.1177/0963721416683996

Zhang, D., Lin, H., & Perrett, D. I. (2020). Apparent Emotional Expression Explains the Effects of Head Posture on Perceived Trustworthiness and Dominance, but a Measure of Facial Width Does Not. *Perception*, *49*(4), 422–438. https://doi.org/10.1177/0301006620909286

# 8. APENDICES

## 8.1. Appendix I: Consent Form (All Studies)

uOttawa

**Consent Form**

## Effects of image modifications on visual recognition of faces, places, and objects

Université d'Ottawa
École de psychologie

University of Ottawa
School of Psychology

**Researchers**

Principal Investigator:

Charles A. Collin[1]
Phone: 613-562-5800, #4296
Email: ccollin@uottawa.ca

Address: University of Ottawa, Ottawa, Ontario, Vanier Building (145 Jean-Jacques Lussier Pvt.), Room 3089B.

Co-Investigators:

Isabelle Boutet 613-562-5800, #2612
Isabelle.Boutet@uottawa.ca

[1] Research supervisor, principal investigator, and lab director.

**Affiliation:** School of Psychology, Faculty of Social Sciences, University of Ottawa.

**Invitation to Participate:** I am invited to participate in the abovementioned research study conducted by the researchers listed above. I am aware that the study is funded by the Canadian Natural Sciences and Engineering Research Council.

**Exclusion Criteria:** I understand that, due to the visual nature of the tasks I will be asked to do, I may not participate if I have a serious eye disease. The most common eye diseases that might preclude participation include glaucoma, retinal detachment,

☎ 613 562-5799
613 562-5147
145 Jean-Jacques-Lussier
Ottawa ON K1N 6N5 Canada

www.uOttawa.ca

macular degeneration and retinitis pigmentosa. Having to wear glasses to see near or far (i.e., myopia, hyperopia and presbyopia) do not constitute eye diseases for this purpose.

**Purpose of the Study:** The purpose of the study is to better understand how we visually recognize and interact with images of objects, faces, and places.

**Participation:** My participation will consist of one session lasting about 30 minutes. During each session, I will be asked to view images on a computer monitor. I will be asked to give responses to the images that I see in accordance with written instructions provided by a computer program. The tasks I may be asked to do are:

1. Indicate whether or not I recognize an image.
2. Indicate if two images presented together are the same or different.
3. Indicate the name associated with an image.
4. Adjust some characteristic of the image (blur, rotation, colour, etc.) until I can recognize it.
5. Indicate whether or not I have seen an image previously.

I will be fully informed as to which particular tasks I will be asked to do before beginning the experiment.

**Risks:** My participation in this study will entail no risks. I understand that I am to discontinue participation as soon as I experience any discomfort.

**Benefits:** My participation in this study will help expand our knowledge of human behaviour and the human brain, particularly in the realm of visual recognition.

**Confidentiality and anonymity:** The information I will share will remain strictly confidential. I understand that the information will be used only for publication and presentation in scholarly media and that my confidentiality will be protected by having the data stored on a password-protected computer that only the researchers have access to. I understand that the data collected might be analyzed and used for student theses. **Anonymity** will be protected in the following manner: Only grouped data will be published and a code number will be used in place of my name on all data reports.

**Conservation of data:** The data collected, which consists of automatically-recorded responses and reaction times to images presented on a computer monitor, will be kept in a secure manner. The data will be stored on computer hard disks on the University of Ottawa campus. Only the researchers will have access to it. It will be kept for a period of 5 years before being securely deleted.

**Compensation:** I understand that if I am participating as part of the University of Ottawa's Integrated System of Participation in Research (ISPR), then I will receive course credit for participating in this research in accordance with the rules of the ISPR. That is, I will receive 1 credit per hour of participation, or fraction thereof. I understand that I may withdraw from the study at any time and nonetheless receive these credits in full.

**Voluntary Participation**: I understand that I am under no obligation to participate and that if I choose to participate, I can withdraw from the study at any time and/or refuse to answer any questions, without suffering any negative consequences. If I choose to withdraw, all my data gathered until the time of withdrawal will be deleted.

**Acceptance:** I, agree to participate in the above research study conducted by the researchers listed at the beginning of this document. I understand that these researchers are affiliated with the School of Psychology, Faculty of Social Sciences, University of Ottawa and that the research is being done under the supervision of Professor Charles A. Collin.

If I have any questions about the study, I may contact the researcher or his supervisor.

If I have any questions regarding the ethical conduct of this study, I may contact the
Protocol Officer for Ethics in Research,
University of Ottawa,
Tabaret Hall, 550 Cumberland Street, Room 154,
Ottawa, ON K1N 6N5.
Tel.: (613) 562-5387
Email: ethics@uottawa.ca

I agree to download a copy of this consent form for my own records.

## 8.2. Appendix II: Demographics Questionnaire

**Demographics Questionnaire**

Please answer the following questions to the best of your ability.

Age:_____

Please select **ALL** that fit with your current gender identity:

- Agender
- Cisgender Female (my gender is the same as the sex I was assigned at birth – I identify as a woman and female)
- Cisgender Male (my gender is the same as the sex I was assigned at birth – I identify as a man and male)
- Genderfluid
- Genderqueer
- Non-binary
- Questioning
- Transgender Female (my gender is different than the sex I was assigned at birth – I am a woman, but I was not assigned female at birth)
- Transgender Male (my gender is different than the sex I was assigned at birth – I am a man, but I was not assigned male at birth)
- Two-Spirit
- You don't have any option that applies to me. I identify as (please specify):_____
- Choose not to respond

What is your ethnicity? Please select **ALL** that apply:

- Arab
- Arab/African
- Asian – East (China, Hong Kong, Japan, Macau, Mongolia, North Korea, South Korea, & Taiwan)
- Asian – South (Bangladesh, Bhutan, India, Pakistan, Nepal, Sri Lanka, Afghanistan, & The Maldives)
- Asian – South East (Brunei, Myanmar, Cambodia, Timor-Leste, Indonesia, Laos, Malaysia, Philippines, Singapore, Thailand, & Vietnam)
- Black – African
- Black – Caribbean
- Black – North American
- First Nations
- Indian – Caribbean
- Indigenous/Aboriginal
- Inuit
- Latin American
- Métis

- Middle Eastern
- Mixed Heritage
- Pacific Islander (American Samoa, Chuuk, Guam, Fiji, Hawaii, Kiribati, Kosrae, Mariana Islands, Marshall Islands, Palau, Papua New Guinea, Pohnpei, Saipan, Samoa, Solomon Islands, Tokelau, Tahiti, Tonga, Vanuatu, & Yap)
- White – North American
- White – European
- I don't know
- You don't have any options that apply to me. My ethnicity is (please specify):_____
- Choose not to respond

Have you lived in Canada for your entire life?

- Yes
- No

How many years have you live in Canada?_____ (Question only asked if they select no to living in Canada their entire life)

If you have lived outside of Canada, please list **ALL** other countries you have lived and for how long:_____ (Question only asked if they select no to living in Canada their entire life)

How many years have you been in university (undergraduate and graduate studies combined)?___

Native (First) Language:

- English
- French
- Other (please specify):_____

Language of Preference for Reading:

- English
- French
- Other (please specify):_____

Do you require corrective lenses for reading on a computer?

- Yes
- No

If so, are you wearing your corrective lenses?

- Yes
- No
- N/A

If participants select Yes to needing corrective lenses, but No to wearing them, they see the following:

Please wear your corrective lenses if you need them for reading on a computer.

Are you know wearing your corrective lenses?

- Yes
- No

## 8.3. Appendix III: Task Instructions

### *8.3.1. Study 1 Task Instructions*

You are about to view several images of people expressing different emotions. Some faces are real, and some are generated by artificial intelligence (AI). For each face, please indicate whether you think it is real or AI-generated. Please note, for both the real and AI-generated faces, you will see images of the same individual with different facial expressions. If you see the same individual multiple times, this is not an indication as to whether the image is real or AI-generated.
To make sure you are paying attention, on some trials, the words "AI-Generated" or "Real" will appear in the image. When you see this, click the corresponding button (i.e., if the words in the image say "AI-Generated", select AI-Generated as your response; if the word in the image says "Real", select Real as your response.)
When you are ready to begin, please click "Begin Experiment".

### *8.3.2. Study 2 Task Instructions*

You are about to view several images of people expressing different emotions. For each face, please rate the emotion expressed on the sliders provided. Please note, if you want to give a rating of "Neutral", all sliders should be set to zero.
To make sure you are paying attention, on some trials, the images will have a number overlaid on top of them. For these trials, set the slider to the number you see on the image.
When you are ready to begin, please click "Begin Experiment".

### *8.3.3. Study 3 Task Instructions: Correct/Incorrect Information Conditions*

You are about to view several images of people expressing different emotions. Some of these images are real photographs and some are AI-generated images. The images will be labelled. For each face, please type up to ten (10) things that you think about them upon first seeing them. These can be your thoughts on their personality, attractiveness, emotional state, likes/dislikes, etc. Anything you feel would be a part of your first impression of them.
To make sure you are paying attention, after some trials, you will be asked to select from several images the one you most recently saw.
When you are ready to begin, please click "Begin Experiment".

### *8.3.4. Study 3 Task Instructions: No Information Condition*

You are about to view several images of people expressing different emotions. For each face, please type up to ten (10) things that you think about them upon first seeing them. These can be your thoughts on their personality, attractiveness, emotional state, likes/dislikes, etc. Anything you feel would be a part of your first impression of them.
To make sure you are paying attention, after some trials, you will be asked to select from several images the one you most recently saw.
When you are ready to begin, please click "Begin Experiment".

## 8.4. Appendix IV: Debriefing Forms

### *8.4.1. Study 2 Debriefing Form*

uOttawa

Université d'Ottawa
École de psychologie

University of Ottawa
School of Psychology

**Debriefing Form**

**Effects of image modifications on visual recognition of faces, places, and objects**

Thank you for your participation in this research study. For this study, it was important that we withhold some information from you. We will describe information that was withheld, why it was important to do so, answer any of your questions, and provide you with the opportunity to decide on whether you would like to have your data included in this study.

**What you should know about this study**

Information was withheld from you regarding the goal of the study. One of the goals of this study is to examine how emotion ratings for AI-generated faces differ from those of real faces. In the set of images you were shown, half of them were real photographs, and the other half were AI-generated. This deception was necessary for two reasons: (1) knowledge that some images were generated by AI may have resulted in you trying to distinguish between the real and AI-generated images, which may have impacted your ratings, and (2) when viewing images in day-to-day life, you will not necessarily know if an image is real or AI-generated when making judgements about it.

**If you have questions**

The main researcher conducting this study is Charles Collin, a professor in the School of Psychology at the University of Ottawa. If you have questions, you may contact Charles Collin at ccollin@uottawa.ca or at 613-562-5800 x4296. If you have any questions or concerns regarding your rights as a research participant in this study, you may contact the Protocol Officer for Ethics in Research at the University of Ottawa at ethics@uottawa.ca or (613) 562-5387.

**Right to withdraw**

You may choose to withdraw the data you provided prior to debriefing, without any penalty of loss of benefits. Please select below if you do, or do not, give permission to have your data included in the study. Selecting an answer below indicates that you have been debriefed, and have had all of your questions answered.

Please download and retain a copy of this form for your own records.

☎ 613 562-5799
613 562-5147

145 Jean-Jacques-Lussier
Ottawa ON K1N 6N5 Canada

www.uOttawa.ca

### *8.4.2. Study 3 Debriefing Form*

**Debriefing Form**

**Effects of image modifications on visual recognition of faces, places, and objects**

uOttawa

Université d'Ottawa
École de psychologie

University of Ottawa
School of Psychology

Thank you for your participation in this research study. For this study, it was important that we withhold some information from you. We will describe information that was withheld, why it was important to do so, answer any of your questions, and provide you with the opportunity to decide on whether you would like to have your data included in this study.

**What you should know about this study**

Information was withheld from you regarding the goal of the study. One of the goals of this study is to examine how social attributions are impacted by knowledge of whether a face is real or AI-generated. One group of participants was shown correctly labelled images of real and AI-generated faces (i.e., if the image was generated by AI it was labelled as such, and if the image was real it was labelled as such); a second group was shown incorrectly labelled images of real and AI-generated faces (i.e., if the image was generated by AI it was labelled as real, and if the image was real it was labelled as AI-generated); a third group was given no information about the images. This deception was necessary to examine the differences between knowledge groups.

**If you have questions**

The main researcher conducting this study is Charles Collin, a professor in the School of Psychology at the University of Ottawa. If you have questions, you may contact Charles Collin at ccollin@uottawa.ca or at 613-562-5800 x4296. If you have any questions or concerns regarding your rights as a research participant in this study, you may contact the Protocol Officer for Ethics in Research at the University of Ottawa at ethics@uottawa.ca or (613) 562-5387.

**Right to withdraw**

You may choose to withdraw the data you provided prior to debriefing, without any penalty of loss of benefits. Please select below if you do, or do not, give permission to have your data included in the study. Selecting an answer below indicates that you have been debriefed, and have had all of your questions answered.

Please download and retain a copy of this form for your own records.

☎ 613 562-5799
613 562-5147

145 Jean-Jacques-Lussier
Ottawa ON K1N 6N5 Canada

www.uOttawa.ca