Considering how machine-learning algorithms (re)produce social biases in generated faces

Matthew Gusdorff[1], Alvin Grissom II[1], Jeova F.S. Rocha Neto[2], Yikang Lin[1], Ryan Trotter[1], & Ryan F. Lei[3]

[1]Haverford College
Department of Computer Science
370 Lancaster Ave
Haverford, PA 19041

[2]Bowdoin College
Department of Computer Science
255 Maine Street
Brunswick, Maine 04011

[3]Haverford College
Department of Psychology
370 Lancaster Ave
Haverford, PA 19041

Word count: 5,757

Corresponding authors: mgusdorff@haverford.edu & rlei1@haverford.edu. The authors declare no conflict of interest.

Acknowledgements. Support for this manuscript comes from an NSF grant awarded to R. Lei and A. Grissom II (CNS-2210142).

Abstract

Advances in computer science–specifically in the development and use of generative machine learning–have provided powerful new tools for psychologists to create synthetic human faces as stimuli, which ultimately provide high-quality photorealistic face images that have many advantages, including reducing typical ethical and privacy concerns and generating face images from minoritized communities that are typically underrepresented in existing face databases. However, there are a number of ways that using machine learning-based face generation and manipulation software can introduce bias into the research process, thus threatening the validity of studies. The present article provides a summary of how one class of recently popular algorithms for generating faces–generative adversarial networks (GANs)--works, how we control GANs, and where biases (with a particular focus on racial biases) emerge throughout these processes. We discuss recommendations for mitigating these biases, as well as how these concepts manifest in similar modern text-to-image algorithms.

Keywords: Face perception, methods; machine learning; bias

Considering biases in the creation and use of computer-generated faces

The face has long held a special place in psychological research. People make all sorts of inferences on the basis of seeing someone's face (Zebrowitz, 2017), and impressions of a person's face can predict a range of important outcomes, from romantic attraction (Rhodes et al., 1999) to voting intentions (Carpinella et al., 2016). Unsurprisingly, psychologists often use static images of faces as stimuli to test a range of research questions (the list of which would take up the entirety of this article). For example, the Chicago Face Database (CFD; Ma et al., 2015), cited over 1,800 times at the time of writing, exemplifies the popularity of such datasets in research.

Although these face databases are very appealing, there are a number of issues with their popularity. For one, face databases are often (if not always) susceptible to contamination by the systematic racial biases that infuse much of social and personality psychology (Cook & Over, 2021; Torrez et al., 2023). Face databases have historically been collections of real people's faces and typically center White people (e.g. Lundqvist et al., 1998), which may reflect psychologists' usual social locations that are majority White (e.g., European countries, university settings). This feature of our methodology is one way that Whiteness becomes centered and seen as a default identity (e.g., Garay & Remedios, 2021; Purdie-Vaughns & Eibach, 2008), often prompting the vexing question of whether and why a White control group might be needed (Syed, 2020).

The lack of attention to representation and diversity in face stimuli can also undermine validity and limit the range of research questions posed. With respect to the range of questions possible, consider research demonstrating that, even beyond broadly perceived race membership, perceived phenotypic cues can shape important consequences, such as capital sentencing (Eberhardt et al., 2006) and self-perceptions among racially minoritized youth (e.g., Rosario et al., 2021). If a face database does not have sufficient diversity in face stimuli for racially minoritized groups, then these questions are not possible to test. Additionally, findings based on a limited stimulus set (e.g., darker-skinned Black people) may not be generalizable or replicable when considering a wider range of stimuli, potentially undermining the validity of the study.

Recently, advances in machine learning have yielded promising new tools to address these issues and match the increasing demand for synthetic faces in psychology (Dawel et al., 2022). Specifically, the creation of artificial photorealistic faces through a class of models known as generative adversarial networks (GANs; Goodfellow et al., 2014) provide high-quality stimuli that are often indistinguishable from real human faces and in fact are rated as more trustworthy than real faces (Nightingale & Farid, 2022). As such, psychologists are increasingly using these tools in their research on basic face perception (e.g., Peterson et al., 2022), as well as focusing on perception of GAN-generated faces in and of themselves (Miller, Steward, et al., 2023; Nightingale & Farid, 2022). These synthetic faces allow researchers to generate new faces that participants have likely never seen (unlike the many studies that use the CFD faces), address ethical concerns of portraying real people as undesirable (e.g., as criminals), and may provide greater experimental control, helping to address stimulus as a random factor in statistical models (Judd et al., 2012).

Despite their promise, GANs, and machine learning more generally, often reify existing social biases (e.g., Jain et al., 2022; Salminen et al., 2020). For example, white computer-generated faces are more often seen as hyper-realistic (i.e., more human than an actual human), while non-White faces are merely photorealistic (Miller, Steward, et al., 2023; Nightingale & Farid, 2022). GANs can even be biased in the expressions on the faces they generate (Muñoz et al., 2023). All such biases can potentially undermine experimental validity. Most psychologists may be unsurprised to learn that these tools and data have biases, but they may not quite understand where and how such biases enter the process. That is, psychologists may be less familiar with the mechanisms underlying GANs, not fully appreciating the parts of the process where bias can affect the generation of artificial faces. The present article seeks to provide a guide for psychologists to better understand these nuances of using GAN-generated faces in order to better understand the risks they can pose to study validity.

A short primer on General Adversarial Networks (GANs)

Computer-generated faces are quite appealing for psychological researchers because they provide a great deal of experimental control (Dawel et al., 2022). Being able to create a face and systematically vary it along a dimension of interest allows a researcher to better control for face dimensions that may exert unintended and unwanted influences. For example, FaceGen (*FaceGen 3D*, 2003), one of the first programs for artificially generating faces, allows a researcher to specify the sex of a face as a base model and then systematically vary a number of other dimensions of interest for that face (e.g., Afrocentric facial features, skin tone, emotional expression). But these earlier programs usually relied on technology which led to limited outputs (i.e., generated stimuli) that were obviously computer generated and not at all photorealistic, consequently leading to study results that deviated from similar ones using strictly human faces (Gaither et al., 2019; Miller, Foo, et al., 2023). Furthermore, teaching a program to know what is “Afrocentric” or “happy” relies on human interference, a process that can exacerbate stereotypes (Otterbacher, 2018).

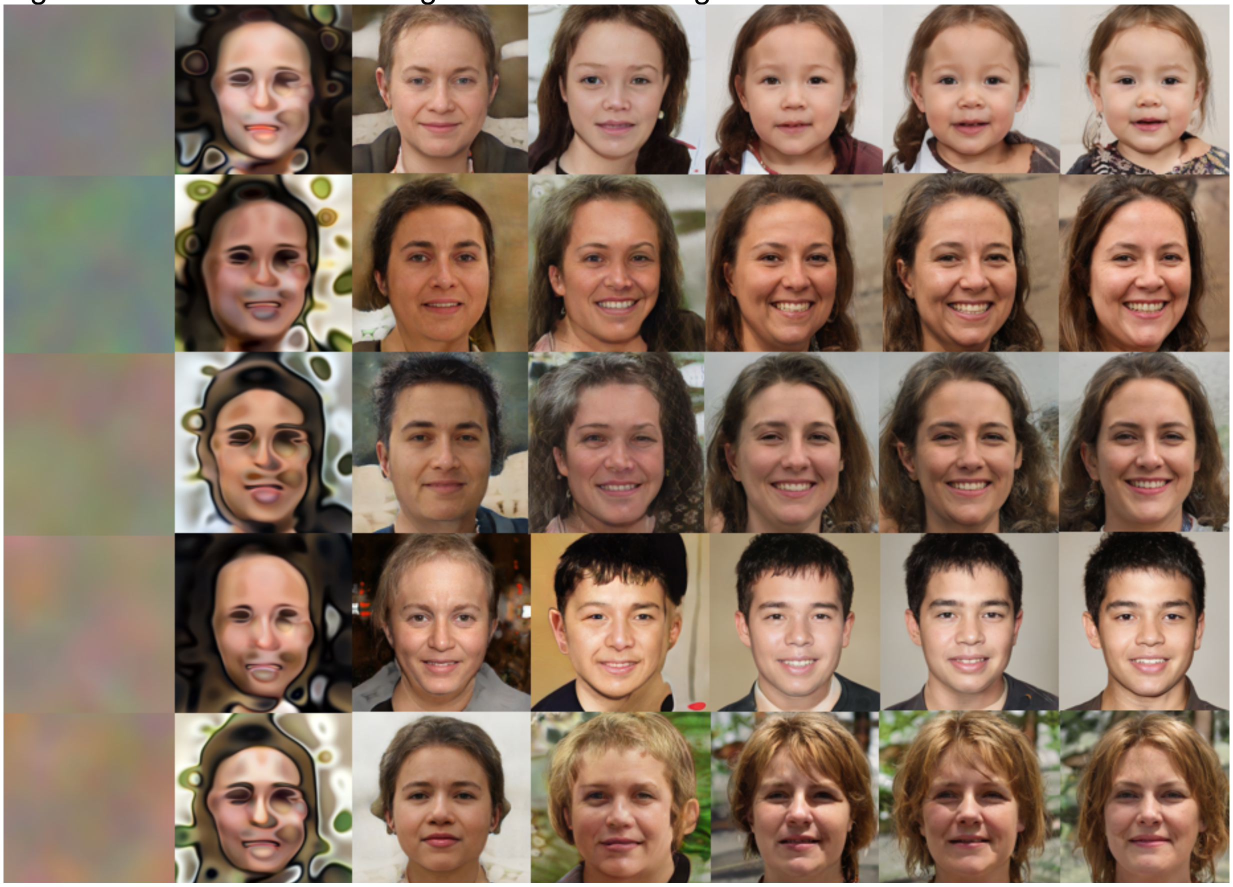
Whereas earlier programs such as FaceGen may have sought to isolate individual facial features that they then converted into a manipulable vector, face generation currently uses much more complex and dynamic modern machine learning algorithms. In general, a machine-learning algorithm begins with an untrained model. By *model*, we mean a statistically predictive function whose parameters are set (i.e., *learned*) from examples, some of which may be familiar to psychologists (e.g., linear or logistic regression models), and some which may be less familiar (e.g., decision trees, neural networks). Most psychologists are versed in using such models to *fit* data, but the same kinds of models can be used to make predictions about what *new* data will look like. This is one way to think of the difference in focus between statistics and machine learning, the former being more concerned with modeling data as it is, and the latter being more concerned with generalizing to new data.

Before training, untrained machine-learning models greatly underperform, producing meaningless outputs (see first column of Fig. 1 for example output). Imagine trying to fit a regression line without feeding it any data. While it may be able to find a best-fit line in principle, without having seen the data, this line will be random, and the model is not useful. But as the algorithm is fed more relevant data points–in this case, faces–the model begins to learn. By *learn*, we mean that the algorithm finds the model parameters that minimize the error, or loss–the loss function is specified depending on the application and model–on the training data. This is akin to how the line begins to rotate and shift to reduce residuals in linear regression, or minimize the loss (i.e., the mean squared error in linear regression).

When the model's error has been minimized to our liking, we say it is *trained*. After a model is trained, it can be used for *inference*. Inference in this case refers to using the model to generate, estimate, or classify new data. In linear regression, this would be providing an input x, from which the model will predict an output $\hat{y}$, an estimation of the true value y, based on the patterns it has learned. This basic process is the same one that more complex algorithms, such as neural networks, use to generate their inferences, which sometimes take the form of numbers, sentences, and images. Note that these data on which inference is done are typically different from the training data and minimizing error on the training data (training error) doesn't necessarily translate to minimizing error when we generalize to these new data points used for inference (generalization error). That is, just because a model learned something during training does not guarantee that knowledge will be applicable when using inputs beyond the training data. For example, a model trained to predict the weather in Australia will not necessarily be useful in Norway.

Face generation software using machine learning follows these general principles. GANs are one of the latest iterations of these models but are made up of two independent networks: the generator and the discriminator. The generator is trained with face images and its objective is to produce novel images of faces that it has not seen before, while the discriminator is trained with half of its images produced by the generator and the other half as real faces. The discriminator outputs a score quantifying its estimate of whether the input face is real or fake, i.e., artificially fabricated by the current version of the generator. The generator learns to create faces that are considered real by the current state of the discriminator. It makes these faces by transforming a vector of random numbers into the face itself (randomness ensures that the model generates a variety of faces). This back-and-forth process is a zero-sum game played by the models, where both seek to win against the other, improving both until the generator generates photorealistic images and the discriminator can detect photorealistic fakes. Once all the training data has passed through the algorithm (known as an epoch), the process occurs over and over again until the researcher is satisfied with the output or a predefined number of epochs has been reached. Figure 1 displays the progression of face generation from untrained (far left column, representing training epoch 0) to completely trained (far right column, representing training epoch 25,000). For broader, more technical surveys of GANs, see Iglesias et al., 2023 and Trevisan de Souza et al., 2023.

Figure 1. A selection of GAN-generated face images.



Note: Each row represents a unique vector to generate faces at various points of the training process represented by each column. In the beginning, the generator generates largely noise, but over time, as the discriminator improved at distinguishing real faces from fakes, the generator's output becomes more convincing, until its faces are realistic.

Sources of bias in artificially generated faces

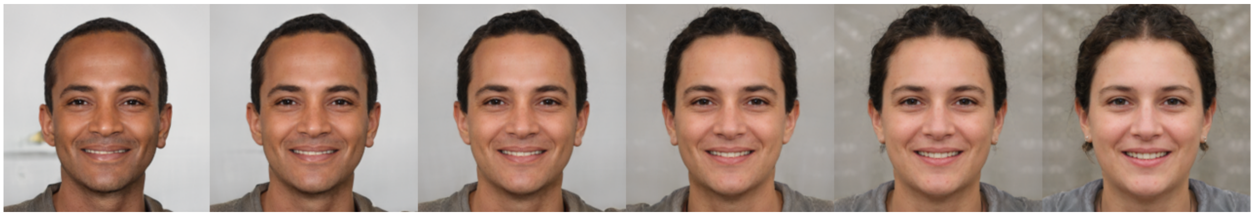
Although GANs represent an exciting new method for generating low-cost, photorealistic faces, there are a number of potential biases that may hamper their usefulness. We highlight these areas of concern not to dissuade psychological researchers from using these images as stimuli, but rather to elucidate potential threats to validity. Although some areas of bias may seem obvious to psychologists, others may not be. Broadly speaking, the most impactful areas where bias can manifest in the creation of artificial faces are in training datasets and model architecture. We discuss each of these in turn, describing how bias can manifest and threaten scientific validity. While we provide some possible avenues for the mitigation of these biases, offering concrete technical solutions is beyond the scope of this survey.

Training datasets as a source of bias

Some psychologists may intuitively believe that machine-learning algorithms learn in the same way that humans learn about the social world–and in some ways, that may be true, e.g., exploiting statistical regularities (Zaadnoordijk et al., 2022). From this perspective, it may be unsurprising that they also encode various biases we see in humans and potentially many others from the data it is exposed to. Bias in training datasets may be the source of bias that psychologists immediately think of–the kinds of images an algorithm observes as it learns is likely to shape what kinds of images the algorithm generates once it has been trained, analogous to how human infants have biases in looking times that are sensitive to their local contexts. For example, psychologists have documented how even by the first few months of life, children are attuned to the racial makeup of their social environment (i.e., inputs) and correspondingly show biases towards people who are statistically more prevalent in their surroundings (Bar-Haim et al., 2006; Hwang & Markson, 2023). GANs operate similarly. The most frequently used training dataset—the Flickr Face-HQ (FFHQ) database—for StyleGAN2, one of the most popular GANs, is overwhelmingly white and slightly more female than male at 52-55%, which is not too far from being representative of the population (Karras et al., 2020; Miller, Steward, et al., 2023; Perera & Patel, 2023): therefore, it may be unsurprising that the modal face that StyleGAN2 generates is a white woman (Salminen et al., 2020).

This bias is illustrated through an important element of NVIDIA's StyleGAN: the *truncation trick* (Karras et al., 2019). Since the algorithm employs randomness in its facial generation, occasionally it can produce faces with odd features (e.g., uneven eyes, missing ears, etc.) or graphical artifacts. The truncation trick solves this problem by limiting the distribution of faces to be closer to the mean image based on a value Ψ (*psi*). Ψ can be a number between 0 and 1, where 0 restricts the output to the mean, and 1 uses no truncation at all. If the model reflects the training data, this mean image will be a white woman; therefore, the more the GAN is restricted via truncation, the whiter and more feminine the output image becomes (see Figure 2). Thus, images of people from minority groups may be excluded when using standard amounts of truncation for increasing image quality and coherency (Maluleke et al., 2022). That is not to say that the truncation trick itself is problematic: it is indeed vital to generating realistic, high-quality faces. However, truncation exposes, even exacerbates, StyleGAN's biases, so one must be careful when employing it.

Figure 2. How truncation shifts an image.



Note: As the *psi* value decreases (from left to right, the image goes from 0.5-0 truncation), the individual generated face image moves closer to the mean image generated.

GAN architecture as a source of bias

While we can control what is in the training data, it is more difficult to control how the GAN learns from it. The mathematical design of a model, its architecture, also plays a key role in the biases it develops. Whereas a linear regression model has a simple architecture consisting of a best-fit line (or hyperplane) with a slope and intercept, more complex models, such as neural networks, are made up of several layers of interconnected “neurons”–in traditional feed-forward neural networks, a neuron is akin to a logistic regression–which collectively contribute to producing an output. These layers are built and organized in such a way that, given proper training, the model as a whole can achieve its objective. However, unlike models such as linear regression, which only have one solution, GANs can have many. Intuitively, this means that there is no *unique* solution for generating a face, or even a given face. Due to the complexity of the architecture and the existence of many possible solutions, seemingly small variations–either in the dataset, the architecture itself, or the training process, which has elements of randomness–can greatly impact not only the quality or correctness of a model’s output, but also the biases it exhibits (Arjovsky et al., 2017). That is, the biases a GAN exhibits come from how the architecture interacts with the data it sees, but since this process is complex (Li et al., 2018), the exact sources of bias in a GAN’s architecture are not deeply understood. Furthermore, since there are many different kinds of bias, which often appear as what we might call side-effects or “pathologies” (Feng et al., 2018; Grissom II et al., 2024), we cannot exhaustively search for every possible manifestation of bias; some biases and their causes may be inscrutable.

However, we can *observe* salient biases in a model. In some recent work, we examined one of NVIDIA’s premier GANs, StyleGAN3-r trained on FFHQ (Karras et al., 2021), to demonstrate racial bias in the discriminator (Grissom II et al., 2024). The discriminator outputs a score indicating its confidence that the image is real (according to the discriminator), with higher scores indicating more confidence. In theory, the model is more likely to generate images more similar to those assigned higher scores than those assigned lower scores. When shown a sample of human-annotated real faces grouped by perceived race, gender, and hair length, the discriminator shows an objective preference for lighter colors, assigning them higher scores, and consequently a preference for Eurocentric and Asiocentric faces over Afrocentric faces (Grissom II et al., 2024). Thus, the GAN is less likely to generate faces with darker skin tones, internalizing the racial bias in FFHQ, its training data, into its architecture.

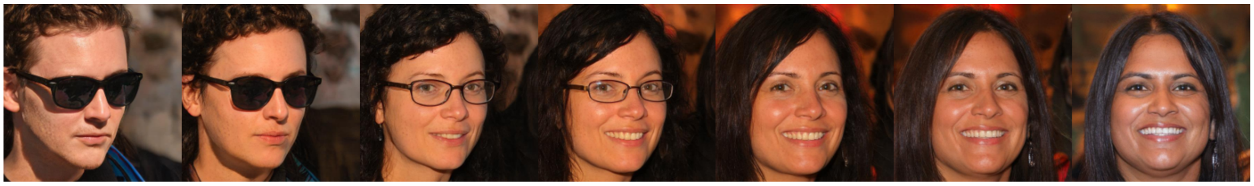
Controlling GANs

Of course, a psychologist looking to generate face images for studies is likely looking to generate stimuli with specific demographic parameters. Although it is possible to get a sufficient number of faces by just having a GAN randomly generate faces until the researcher is satisfied, this approach is inefficient at best. Instead, a psychologist would likely prefer to press a series of buttons to get, for example, Black, White, and Asian male and female faces, roughly 20 years old in appearance. One avenue for achieving control is designing a GAN to consider annotated data–that is, images that have been rated by humans on any number of desired dimensions (e.g., gender, race, skin tone, etc.). One such GAN is GAN-Control (Shoshan et al., 2021). GAN-Control requires annotations for each training image. By considering the face image and its features in tandem with annotations, the GAN can build connections between the pixels it sees and the provided labels. When GAN-Control generates new images, it no longer receives solely a random number, but also requested features that it can then put into a face.

Another approach to controlling GANs is *latent space editing*. Every GAN learns a *latent space*, a part of the model that specifies the possible images the GAN's generator can generate. The generator "interprets" the numbers in this latent space to generate images. The latent space can be used to one end: controlling the GAN. A common application of editing in the latent space is face morphing: that is, you can generate two faces, and, in this mathematical latent space, interpolate along the line that connects the two images by mixing their latent representations, effectively morphing them.

This idea can be extended to general *feature directions*. Certain features of faces are grouped in the latent space. Therefore, instead of walking a line between two images to morph them, one can gradually introduce a feature into an image by walking in the direction of that feature in the latent space (Shen et al., 2020). Take, for example, Figure 3, where two faces are morphed. While these faces vary along many dimensions, one distinct difference is the presence of glasses. As the faces are morphed, we see the glasses (dis)appear. This implies that glasses are associated with a certain feature direction which we can navigate in the latent space. In fact, it is possible to find this direction precisely, as demonstrated by Shen et al., 2020. The ability to generalize feature directions allows a vast amount of customizability in GAN-generated faces, ranging from wearing glasses to age, gender, and more.

Figure 3. An example of how interpolation works



Note: Interpolation involves morphing faces into a unique image from two distinct starting images. Drawn from https://facemorph.me/?from_value=ella&to_value=lillian.

Neither of these approaches are without its flaws, however. In the case of annotating images for GAN-Control or similar algorithms, an additional human variable is injected into the training process. Annotation by human raters can exacerbate biases even if the underlying training dataset is perfectly balanced. Take, for example, a situation where a researcher wanted to curate a balanced dataset of faces that had equal numbers of White, Black, Latinx and Asian faces. If some of the faces in the dataset were ambiguous for any reason (e.g., reflects someone who is biracial, lighting biases, skin tone perception), then raters could miscategorize the face. This kind of miscategorization can happen depending on any number of combinations of perceptual features that do or do not align with perceivers' notions of what it means to belong to a given racial category (Maddox, 2004; Nicolas et al., 2019). As a result of the miscategorization, the GAN is then trained on an unbalanced dataset, despite the researcher's best intentions to curate a balanced dataset. Additionally, even with excellently annotated training data, there is no guarantee that the GAN can draw meaningful connections between our human labels and its mathematical model of faces, possibly impacting quality and label correctness.

Latent space editing suffers from a similar issue. While interpolating over an existing latent space does not impose our human labels directly on the GAN's training, biases still seep in while finding feature directions. In order to find feature directions, we must use other machine-learning techniques that require annotations: we cannot simply "ask" the GAN how to find certain features in the latent space, because it has no notion of human labels. Instead, we must apply labels to the latent space *after* training (Shen et al., 2020). In this sense, using GAN-Control or similar methods forces the GAN to learn labels alongside faces so it can generate labeled faces on its own, whereas the editing approach interprets and labels the faces the GAN already knows how to make.

An interesting use of the latent space is through self-conditioning during training (Liu et al., 2022). A GAN can be programmed so that during training, it can consider what parts of the latent space–what kinds of faces–are being undergenerated and compare them to a set of labels, allowing the model to determine whether faces with certain attributes–say, with darker skin tones–are not being properly represented in the GAN's output distribution. The GAN then adjusts its parameters to compensate for this disparity. This allows for more balanced outputs which are easily navigated with previously discussed latent space editing methods. Unfortunately, although this approach can better produce minority labels, the variety within these labels can be limited (Liu et al., 2022).

Regardless of how we interact with the GAN, we cannot eliminate every source of bias or completely control every aspect of GANs. Faces are extremely complex, and while the latent space is high dimensional, we cannot guarantee that the features it learns are free of bias or even salient to humans. Thus, it is likely to entangle features, such as long hair and earrings, making meaningful and discrete feature directions altogether difficult to find. In this way, the correlations between features that GANs learn can amplify stereotypic associations. For example, although long hair and earrings are typically correlated because women often have both, they are not intrinsically part of how we want to represent a woman. A human might be able to (relatively easily) imagine a man with long hair or earrings, but a GAN that has learned these correlated features may not represent these features orthogonally and generate more atypical exemplars, exacerbating stereotypic associations. To address these problems, one might be tempted to train separate GANs for the features they wish to isolate. Not only is this process resource intensive, but it also introduces more human biases to the process, baking in the researcher's notions of specific groups or facial features into the data and the models' creation. In summary, although we may find similarities between how a GAN interprets faces and forms biases and our own processes of doing so, they are altogether different; it is important to be wary that the learned representations of a machine-learning model are not necessarily commensurate with those of a human.

Approaches to mitigate bias

While we cannot *perfect* the process of generating and controlling GANs, there are nonetheless some approaches that can help attenuate the impact of various biases on the creation of GAN-generated faces. Given the technical nature of these issues, most solutions will require some level of technological knowledge. To address some issues of bias in the training image set, one possibility, which requires great expertise, is to train a new GAN algorithm on a new, more diverse image set. This is possible with newer face databases such as FairFace (Kärkkäinen & Joo, 2019), which has equal representation of seven different racial and ethnic groups (though Grissom II et al., 2024 show that even basic color bias is not explicable purely due to the training data). Other algorithms trained on the FairFace dataset show improved classification abilities of racially minoritized faces (Kärkkäinen & Joo, 2019), and Maluleke et al., 2022 have shown that StyleGAN approximately reproduces the racial distribution of its training data, so it stands to reason that a GAN trained on a more diverse image set such as FairFace would similarly be more likely to generate a more diverse set of faces.

Another facet of training datasets that can be important to consider when thinking of how to mitigate bias is how to minimize the biases of raters. That is, to the extent that people want to generate faces along particular demographic dimensions (e.g., race, gender, age), this requires that the images be pre-rated, so the algorithm can learn which vectors correspond with which identity dimensions. Of course, because the algorithm is learning these vectors based on human input, it will codify any biases that human raters bring to the table. Thus, the strategies that psychological researchers might use to mitigate bias in their typical studies should also be put into use when gathering image ratings: high inter-rater reliability, having multiple raters, ensuring rater diversity with concern for in-group and out-group effects, verifying raters' ability to recognize faces, etc. all remain vital in this context; (Ramon et al., 2019; Wan et al., 2017). Although this can be tedious, it is preferable to the alternative of having face images algorithmically rated, since having an algorithm rate the images will almost certainly reify existing biases.

However, training a GAN on a more diverse image set, even with high quality human raters, may not be sufficient because of biases introduced by rater biases as well as the GAN architecture itself (Grissom II et al., 2024). Here, researchers could manually intervene more directly in order to mitigate potential biases. Recall that GANs rely on the truncation trick, where the algorithm limits the potential range of faces generated. Although this trick is used to prevent the generation of faces that have highly unlikely facial features (e.g., asymmetrical eyes), it can also inadvertently lead to the systematic exclusion of racially minoritized faces (Maluleke et al., 2022). That is, if the discriminator part of GAN has decided that Black faces are unlikely to be "real" faces, then they may be unlikely to be generated in the first place, because the GAN's generator has learned that these faces are unlikely to successfully trick the discriminator. It is important to note, however, that this mode of manual intervention would require removing images with extreme artifacting or other problems by hand, possibly introducing a new form of bias just as manually selecting participants instead of taking a random sample could bias a typical study.

These two methods are likely to be most effective when used in conjunction, potentially mitigating biases in both the training set and the model architecture simultaneously. On their own, each may prove insufficient. Training on a diverse dataset may nonetheless result in the GAN algorithm learning some other color bias, ultimately still resulting in an imbalanced face stimulus set. And merely increasing the Ψ value to decrease truncation might merely reduce the quality of the image without addressing any of the biases. However, when both solutions are used in conjunction, then the resulting generated set of faces may be more diverse and also still fairly photorealistic.

In addition to these more immediate approaches that address existing issues in how GANs generate these faces, we can also consider re-structuring the algorithm so that it avoids these biases in the first place. That is, part of what happens with GANs is that they learn correlations between different features in order to generate these photorealistic faces; thus, long hair becomes associated with more feminine faces, wider noses with darker skin tones, etc. These learned correlations make it difficult to generate stimuli that are counterstereotypical–for example, generating images of men with long hair. It may be helpful to reconfigure the GAN learning algorithm so that features are learned as orthogonal, disentangled elements, and pair the GAN with a control mechanism, e.g., as done with GAN-Control (Shoshan et al., 2021), so researchers can better control the combinations of facial features that appear in their generated sample. Doing so would be particularly important in capturing the full diversity of human experience. We think it's important to note here that this is likely beyond the realm of psychologists: reducing learned, internal biases is a task that the creators of these algorithms still struggle with.

One other consideration does not have much to do with the GAN at all, but rather in the disclosure of these faces' authenticity. Although GAN-generated photorealistic faces are found to be trustworthy (Nightingale & Farid, 2022), humans tend to react negatively to these synthetic faces when they are discernibly artificial (e.g., using FaceGen faces; Balas & Pacella, 2017) or even merely when they are told the faces are fake (Liefooghe et al., 2023). This phenomenon raises another question for the implementation of GANs in psychological studies: specifically, that their inclusion may depend on deception, which may not be the intentions of the researcher, unless the biases arising from said disclosure are not of concern to the study. GANs also have the potential for “leakage”. That is, they may reproduce their training data in part or in whole (Tinsley et al., 2021). To account for this occurrence, researchers should look into the sources and copyright status of any model’s training data.

Generalizability

Thus far, we have described the issue and potential solutions as they pertain to a specific GAN model: StyleGAN. GANs introduced the first wave of truly photorealistic facial generation: their realism, consistency, accessibility, and speed make them excellent tools for psychologists. But these models have evolved to feature text-to-image customization in recent years by modifying the existing GAN methodology (e.g., GigaGAN; Kang et al., 2023), or through new approaches (e.g., diffusion models) trained on datasets such as LAION-5B, which is made up of over five billion images and text pairs (Schuhmann et al., 2022). Models such as DALL-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022) have grown in popularity and have brought many new innovations to generative machine learning. Once trained, such models can associate words with images, requiring only a prompt to generate something unique and customizable. For psychologists, these models can be asked to generate faces of specified demographics, poses, expressions, and more, circumventing much of the difficulty involved with controlling GANs. Despite this impactful utility, these models have shortcomings of their own.

Like ordinary GANs, text-to-image models have a random element to them to ensure diversity in output images. Therefore, when engineering a prompt, the user must be very specific to establish a controlled baseline: requesting certain backgrounds, clothes, age ranges, etc. must be considered, potentially making the task more time-intensive than using a GAN. Moreover, many text-to-image models can generate images using multiple styles (e.g., photorealistic and paintings), adding an additional layer to consider for researchers.

Despite their different architectures, text-to-image GANs and diffusion models also have analogous biases to the GANs we discuss here. When asked to generate an image with the prompt “a person”, popular models (such as DALL-E and Stable Diffusion) will output white people and men far more often than not (Naik & Nushi, 2023). The caption-based element of these models adds a new avenue for bias to enter as well. When prompted with occupations such as “CEO”, “doctor”, or “computer programmer”, these models will output mostly men, but when prompted with “nurse” or “housekeeper”, the results are almost entirely women (Naik & Nushi, 2023), propagating known biases in word embeddings (Bolukbasi et al., 2016) that are a part of their architecture. Even if someone were to specify the gender (e.g., woman CEO), a well-engineered prompt is subject to biases that go beyond commonly incorporated demographics: clothing, location (i.e., background information), and other factors can impact the face itself (Chinchure et al., 2023).

With these flaws in mind, we encourage psychologists to try a variety of these new models and experiment with inputs as detailed, reliable prompts can yield a high degree of quality and control, especially as these models continue to improve and researchers find new ways to identify and mitigate their biases. Given how quickly they have already advanced, text-to-image models may be more accessible, as creating a good prompt can be an easier process than using StyleGAN (depending on the expertise of the researcher or the research goals). For a deeper dive into how these models work, see Yang et al., 2023.

In general, computer-generated faces represent a major potential resource for psychologists who are interested in developing high-quality, easy to obtain face image stimuli for use in studies. However, using them without careful consideration of how these stimuli either match or do not match the populations they purport to represent is a major concern for psychological scientists. We suggest that being cognizant of the potential biases that may influence each step of the artificial face-generation process is critical to ensuring we do not replicate the very biases we aim to study.

Works Cited

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein generative adversarial networks*. 214–223.

Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior*, *77*, 240–248. https://doi.org/10.1016/j.chb.2017.08.045

Bar-Haim, Y., Ziv, T., Lamy, D., & Hodes, R. M. (2006). Nature and Nurture in Own-Race Face Processing. *Psychological Science*, *17*(2), 159–163. https://doi.org/10.1111/j.1467-9280.2006.01679.x

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Carpinella, C. M., Hehman, E., Freeman, J. B., & Johnson, K. L. (2016). The Gendered Face of Partisan Politics: Consequences of Facial Sex Typicality for Vote Choice. *Political Communication*, *33*(1), 21–38. https://doi.org/10.1080/10584609.2014.958260

Chinchure, A., Shukla, P., Bhatt, G., Salij, K., Hosanagar, K., Sigal, L., & Turk, M. (2023). *TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models* (No. arXiv:2312.01261). arXiv. https://doi.org/10.48550/arXiv.2312.01261

Cook, R., & Over, H. (2021). Why is the literature on first impressions so focused on White faces? *Royal Society Open Science*, *8*(9), 211146. https://doi.org/10.1098/rsos.211146

Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2022). A systematic survey of face stimuli used in psychological research 2000–2020. *Behavior Research Methods*, *54*(4), 1889–1901. https://doi.org/10.3758/s13428-021-01705-3

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-

Sentencing Outcomes. *Psychological Science*, *17*(5), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

FaceGen 3D. (2003). [Computer software]. Singular Inversions. https://facegen.com/3dprint_demo.htm

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., & Boyd-Graber, J. (2018). Pathologies of Neural Models Make Interpretations Difficult. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3719–3728. https://doi.org/10.18653/v1/D18-1407

Gaither, S. E., Chen, J. M., Pauker, K., & Sommers, S. R. (2019). At face value: Psychological outcomes differ for real vs. computer-generated multiracial faces. *The Journal of Social Psychology*, *159*(5), 592–610. https://doi.org/10.1080/00224545.2018.1538929

Garay, M. M., & Remedios, J. D. (2021). A review of White-centering practices in multiracial research in social psychology. *Social and Personality Psychology Compass*, *15*(10), e12642. https://doi.org/10.1111/spc3.12642

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (No. arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Grissom II, A., Lei, R. F., Gusdorff, M., Neto, J. F. S. R., Lin, B., & Trotter, R. (2024). *Examining Pathological Bias in a Generative Adversarial Network Discriminator: A Case Study on a StyleGAN3 Model* (No. arXiv:2402.09786). arXiv. https://doi.org/10.48550/arXiv.2402.09786

Hwang, H. G., & Markson, L. (2023). Black and White children’s race-based information endorsement and teacher preference: Effects of school and neighborhood racial demographics. *Developmental Psychology*, *59*(5), 893–907. https://doi.org/10.1037/dev0001507

Iglesias, G., Talavera, E., & Díaz-Álvarez, A. (2023). A survey on GANs for computer vision: Recent research, analysis and taxonomy. *Computer Science Review*, *48*, 100553. https://doi.org/10.1016/j.cosrev.2023.100553

Jain, N., Olmo, A., Sengupta, S., Manikonda, L., & Kambhampati, S. (2022). Imperfect ImaGANation: Implications of GANs exacerbating biases on facial data augmentation and snapchat face lenses. *Artificial Intelligence*, *304*, 103652. https://doi.org/10.1016/j.artint.2021.103652

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., & Park, T. (2023). *Scaling Up GANs for Text-to-Image Synthesis*. 10124–10134. https://openaccess.thecvf.com/content/CVPR2023/html/Kang_Scaling_Up_GANs_for_Text-to-Image_Synthesis_CVPR_2023_paper.html

Kärkkäinen, K., & Joo, J. (2019). *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age* (No. arXiv:1908.04913). arXiv. https://doi.org/10.48550/arXiv.1908.04913

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, *34*, 852–863.

Karras, T., Laine, S., & Aila, T. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks* (No. arXiv:1812.04948). arXiv. http://arxiv.org/abs/1812.04948

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of stylegan*. 8110–8119.

Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the Loss Landscape of Neural Nets. *Advances in Neural Information Processing Systems*, *31*.

https://proceedings.neurips.cc/paper_files/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html

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, Y., Gal, R., H. Bermano, A., Chen, B., & Cohen-Or, D. (2022). Self-Conditioned GANs for Image Editing. *ACM SIGGRAPH 2022 Conference Proceedings*, 1–9. https://doi.org/10.1145/3528233.3530698

Lundqvist, D., Flykt, A., & Öhman, A. (1998). Karolinska Directed Emotional Faces. *PsycTESTS Dataset*, *91*, 630. https://doi.org/10.1037/t27732-000

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Maddox, K. B. (2004). Perspectives on racial phenotypicality bias. *Personality and Social Psychology Review*, *8*(4), 383–401.

Maluleke, V. H., Thakkar, N., Brooks, T., Weber, E., Darrell, T., Efros, A. A., Kanazawa, A., & Guillory, D. (2022). *Studying Bias in GANs through the Lens of Race* (No. arXiv:2209.02836). arXiv. http://arxiv.org/abs/2209.02836

Miller, E. J., Foo, Y. Z., Mewton, P., & Dawel, A. (2023). How do people respond to computer-generated versus human faces? A systematic review and meta-analyses. *Computers in Human Behavior Reports*, *10*, 100283. https://doi.org/10.1016/j.chbr.2023.100283

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

Muñoz, C., Zannone, S., Mohammed, U., & Koshiyama, A. (2023). *Uncovering Bias in Face Generation Models* (No. arXiv:2302.11562). arXiv. https://doi.org/10.48550/arXiv.2302.11562

Naik, R., & Nushi, B. (2023). Social Biases through the Text-to-Image Generation Lens. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 786–808. https://doi.org/10.1145/3600211.3604711

Nicolas, G., Skinner, A. L., & Dickter, C. L. (2019). Other Than the Sum: Hispanic and Middle Eastern Categorizations of Black–White Mixed-Race Faces. *Social Psychological and Personality Science*, *10*(4), 532–541. https://doi.org/10.1177/1948550618769591

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.

Otterbacher, J. (2018). Social Cues, Social Biases: Stereotypes in Annotations on People Images. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *6*, 136–144. https://doi.org/10.1609/hcomp.v6i1.13320

Perera, M. V., & Patel, V. M. (2023). Analyzing Bias in Diffusion-based Face Generation Models. *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10. https://doi.org/10.1109/IJCB57857.2023.10449200

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), e2115228119. https://doi.org/10.1073/pnas.2115228119

Purdie-Vaughns, V., & Eibach, R. P. (2008). Intersectional Invisibility: The Distinctive Advantages and Disadvantages of Multiple Subordinate-Group Identities. *Sex Roles*, *59*(5), 377–391. https://doi.org/10.1007/s11199-008-9424-4

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *Proceedings of the 38th International*

Conference on Machine Learning, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, *110*(3), 461–479. https://doi.org/10.1111/bjop.12368

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science*, *10*(1), 52–58.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-Resolution Image Synthesis With Latent Diffusion Models*. 10684–10695. https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html

Rosario, R. J., Minor, I., & Rogers, L. O. (2021). "Oh, You're Pretty for a Dark-Skinned Girl": Black Adolescent Girls' Identities and Resistance to Colorism. *Journal of Adolescent Research*, *36*(5), 501–534. https://doi.org/10.1177/07435584211028218

Salminen, J., Jung, S., Chowdhury, S., & Jansen, B. J. (2020). Analyzing Demographic Bias in Artificially Generated Facial Pictures. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3334480.3382791

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, *35*, 25278–25294.

Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020). *Interpreting the latent space of gans for semantic face editing*. 9243–9252.

Shoshan, A., Bhonker, N., Kviatkovsky, I., & Medioni, G. (2021). *Gan-control: Explicitly controllable gans*. 14083–14093.

Syed, M. (2020). *Whither the “White Control Group“? On the Benefits of a Comparative Ethnic Minority Psychology*. OSF. https://doi.org/10.31234/osf.io/n4p73

Tinsley, P., Czajka, A., & Flynn, P. (2021). *This Face Does Not Exist... But It Might Be Yours! Identity Leakage in Generative Models*. 1320–1328. https://openaccess.thecvf.com/content/WACV2021/html/Tinsley_This_Face_Does_Not_Exist..._But_It_Might_Be_Yours_WACV_2021_paper.html

Torrez, B., Hudson, S. T. J., & Dupree, C. H. (2023). Racial equity in social psychological science: A guide for scholars, institutions, and the field. *Social and Personality Psychology Compass*, *17*(1), e12720. https://doi.org/10.1111/spc3.12720

Trevisan de Souza, V. L., Marques, B. A. D., Batagelo, H. C., & Gois, J. P. (2023). A review on Generative Adversarial Networks for image generation. *Computers & Graphics*, *114*, 13–25. https://doi.org/10.1016/j.cag.2023.05.010

Wan, L., Crookes, K., Dawel, A., Pidcock, M., Hall, A., & McKone, E. (2017). Face-blind for other-race faces: Individual differences in other-race recognition impairments. *Journal of Experimental Psychology: General*, *146*(1), 102–122. https://doi.org/10.1037/xge0000249

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys*, *56*(4), 105:1-105:39. https://doi.org/10.1145/3626235

Zaadnoordijk, L., Besold, T. R., & Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, *4*(6), Article 6. https://doi.org/10.1038/s42256-022-00488-2

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