



A Study on Generative Adversarial Networks Exacerbating Social Data Bias

Thesis by Niharika Jain

Chair: Dr. Subbarao Kambhampati

Committee Members: Dr. Huan Liu and Dr. Lydia Manikonda



Forbes

6,459 views | Nov 5, 2018, 12:31am

Does Synthetic Data Hold The Secret To Artificial Intelligence?

Bernard Marr Contributor
Enterprise & Cloud

Could [synthetic data](#) be the solution to rapidly train artificial intelligence (AI) algorithms? There are advantages and disadvantages to synthetic data; however, many technology experts believe that synthetic data is the key to democratizing machine learning and to accelerate testing and adoption of artificial intelligence algorithms into our daily lives.

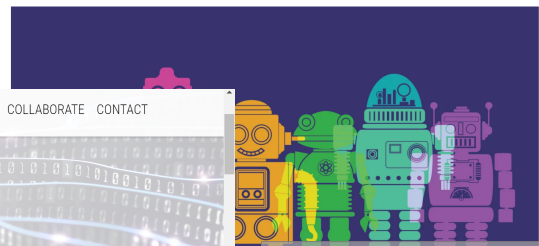
<https://www.forbes.com/sites/bernardmarr/synthetic-data-hold-the-secret-to-artificial-intelligence/>

TC

Startups
Apps
Gadgets
Events
Videos
Crunchbase
More
Search

Deep learning with synthetic data will democratize the tech industry

Evan Nisselson @nisselson / 6 months ago



Disrupt Berlin 2018
48 Hour Pass Sale
Berlin
Nov 29 - 30
[Get your pass now](#)

<https://techcrunch.com/2018/05/11/deep-learning-with-synthetic-data-will-democratize-the-tech-industry/>

SYNTHEA™

HOME ABOUT TECHNOLOGY COLLABORATE CONTACT

Synthetic Patient Generation

Realistic Health Data
No Cost, No Restrictions

SYNTHEA EMPOWERS DATA-DRIVEN HEALTH IT



MIT News

Browse or Search

Artificial data give the same results as real data — without compromising privacy

New approach can help organizations scale their data science efforts with artificial data and crowdsourcing.

Stefanie Koperniak | Institute for Data, Systems, and Society
March 3, 2017

Although data scientists can gain great insights from large data sets — and can ultimately use these insights to tackle major challenges — accomplishing this is much easier said than done. Many such efforts are stymied from the outset, as privacy concerns make it difficult for scientists to access the data they would like to work with.

In a [paper](#) presented at the IEEE International Conference on Data Science and Advanced Analytics, members of the Data to AI Lab at the MIT Laboratory for Information and Decision Systems (LIDS) Kalyan Veeramachaneni, principal research scientist in LIDS and the Institute for Data, Systems, and Society (IDSS) and co-authors Neha Patki and Roy Wedge describe a machine learning system that automatically creates synthetic data — with the goal of enabling data science efforts that, due to a lack of access to real data, may have otherwise not left the ground. While the use of authentic data can cause significant privacy concerns, this synthetic data is completely different from that produced by real users — but can still be used to develop and test data science algorithms and models.

RELATED

- Data to AI Lab
- Institute for Data, Systems, and Society
- Laboratory for Information and Decision Systems
- School of Engineering

<http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303>

Machine learning practitioners have celebrated Generative Adversarial Networks as an economical technique to augment their training sets for data-hungry models when acquiring real data is expensive or infeasible.

It's not clear that they realize the dangers of this approach!

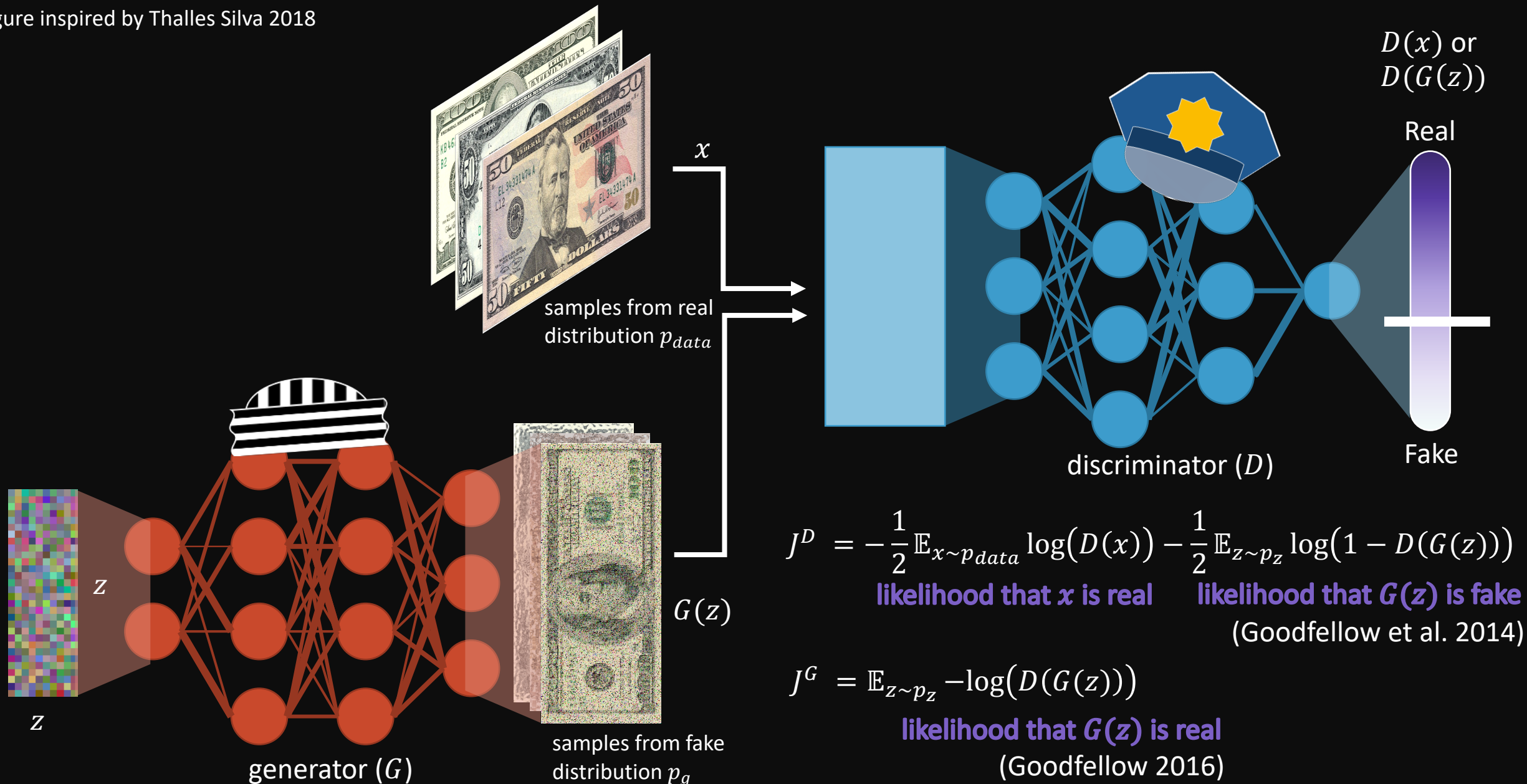
data augmentation

If GANs worked perfectly, they would capture the distribution of the data, and thus capture any biases within it.

GANs have a failure mode which causes them to *exacerbate* bias.

Generative Adversarial Networks: counterfeiter and cop

Figure inspired by Thalles Silva 2018



carpedm20 / DCGAN-tensorflow

Watch 244 Star 6.5k Fork 2.6k

Deep Convolutional Generative Adversarial Networks (DCGAN)

A tensorflow implementation of "Deep Convolutional Generative Adversarial Networks" <http://carpedm20.github.io/faces/>

tensorflow dcgan gan generative-model

299 commits 2 branches 0 packages 0 releases 40 contributors MIT

github.com/carpedm20/DCGAN-tensorflow

(Radford, Metz, and Chintala 2015)

GANs are explosively popular, in part, because scalable models are readily available off-the-shelf.

github.com/junyanz/pytorch-CycleGAN-and-pix2pix

(Zhu et al. 2017)

junyanz / pytorch-CycleGAN-and-pix2pix

Watch 312 Star 11.7k Fork 3.4k

Cycle-Consistent Adversarial Networks (CycleGAN)

Image-to-Image Translation in PyTorch

pytorch gan cyclegan pix2pix deep-learning computer-vision computer-graphics image-manipulation image-generation generative-adversarial-network gans

445 commits 3 branches 0 packages 0 releases 45 contributors View license

A grid of 200 small portrait images of faces, arranged in 10 rows and 20 columns. The faces are generated by a Generative Adversarial Network (GAN) trained on a dataset of engineering professors. The faces exhibit a variety of features, including different hair colors, styles, and colors, as well as various expressions and accessories like glasses. The overall appearance is that of a diverse set of individuals, though the features are somewhat constrained by the training data.

What do these images have in common?

These are GAN-generated faces, trained on a dataset of engineering professors.

hypothesis:

when a feature is biased in the training set, a GAN amplifies the biases along that dimension in its generated distribution

all biases are equal,
but some are more equal than others.

This hypothesis makes a blanket claim about GANs indiscriminately picking up all types of biases that can exist in the data. For facial images, these biased features could be lighting, facial expression, accessories, or hairstyle.

We only aim to bring attention to exacerbation of *sensitive* features: social characteristics that have been historically discriminated against. This work investigates bias over race and gender.

hypothesis:

when a feature is biased in the training set, a GAN amplifies the biases along that dimension in its generated distribution

for facial datasets, these datasets are often skewed along race and gender, so GANs exacerbate sensitive social biases

don't try this at home!

Using photos to measure human characteristics has a complicated and dangerous history: in the 19th century, “photography helped to animate—and lend a ‘scientific’ veneer to—various forms of phrenology, physiognomy, and eugenics.” (Crawford and Paglen 2019)

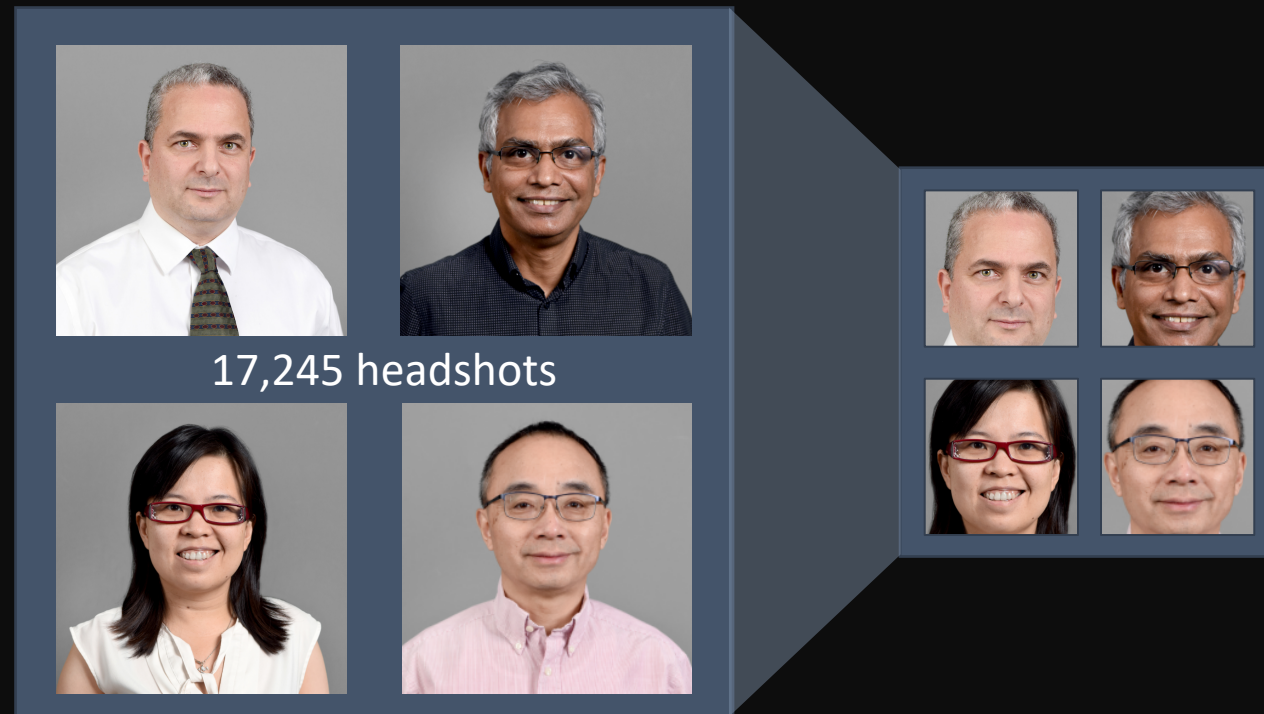
Neither gender nor race can be ascertained from appearance. We use human annotators to classify masculinity of features and lightness of skin color as a crude metric of gender and race to illustrate our argument.

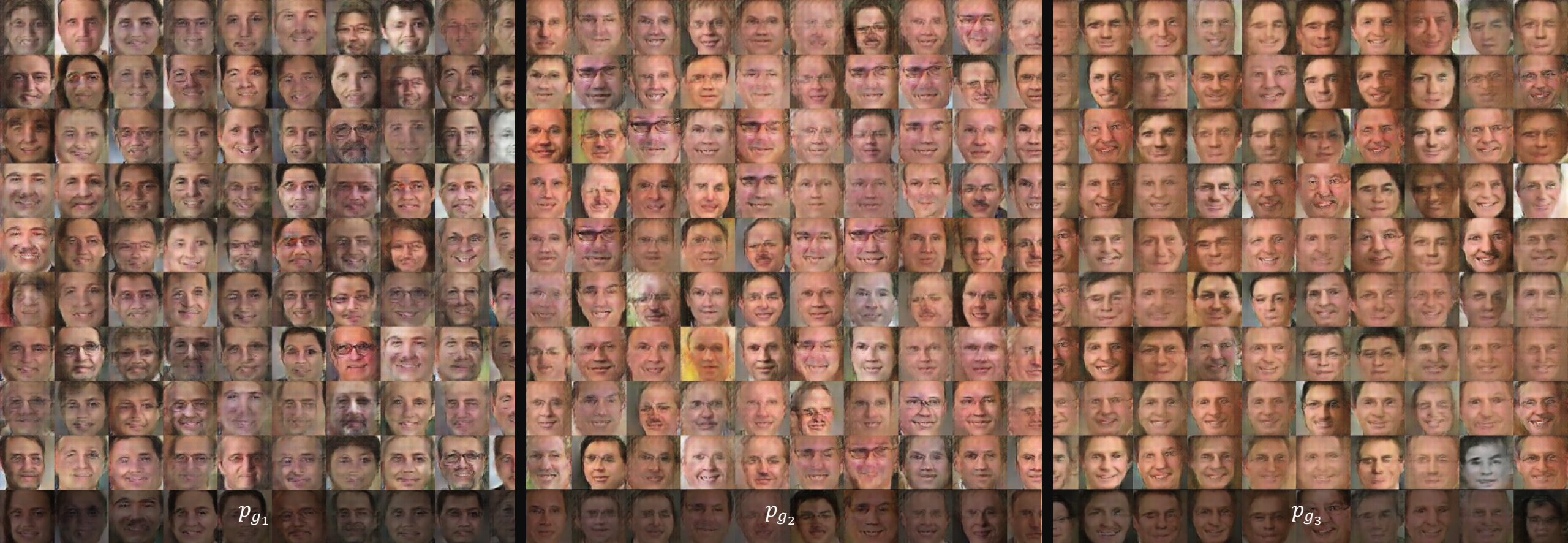
This work is not advocating for the use of facial data in machine learning applications. We create a hypothetical experiment using data with easily-detectable biases to tell a cautionary tale about the shortcomings of this approach.

imagining an engineer

if we train a GAN to imagine faces of US university engineering professors, will it skew the new data toward white males?

We scrape from engineering faculty directories from 47 universities on the U.S. News “Best Engineering Schools” list, remove all noisy images, and crop to the face.





DCGAN trained on three random initializations

To measure the distributions in their diversity along gender and race, we ask humans on Amazon Mechanical Turk to annotate the images.

For each task, we ask master Turkers to annotate 50 images:

T1a gender on professor images randomly sampled from p_{data}

T1b gender on DCGAN-generated images randomly sampled from p_g

T2a race on professor images randomly sampled from p_{data}

T2b race on DCGAN-generated images randomly sampled from p_g

evaluation



For each image, select the most appropriate description:

- face has mostly masculine features
- face has mostly feminine features
- neither of the above is true

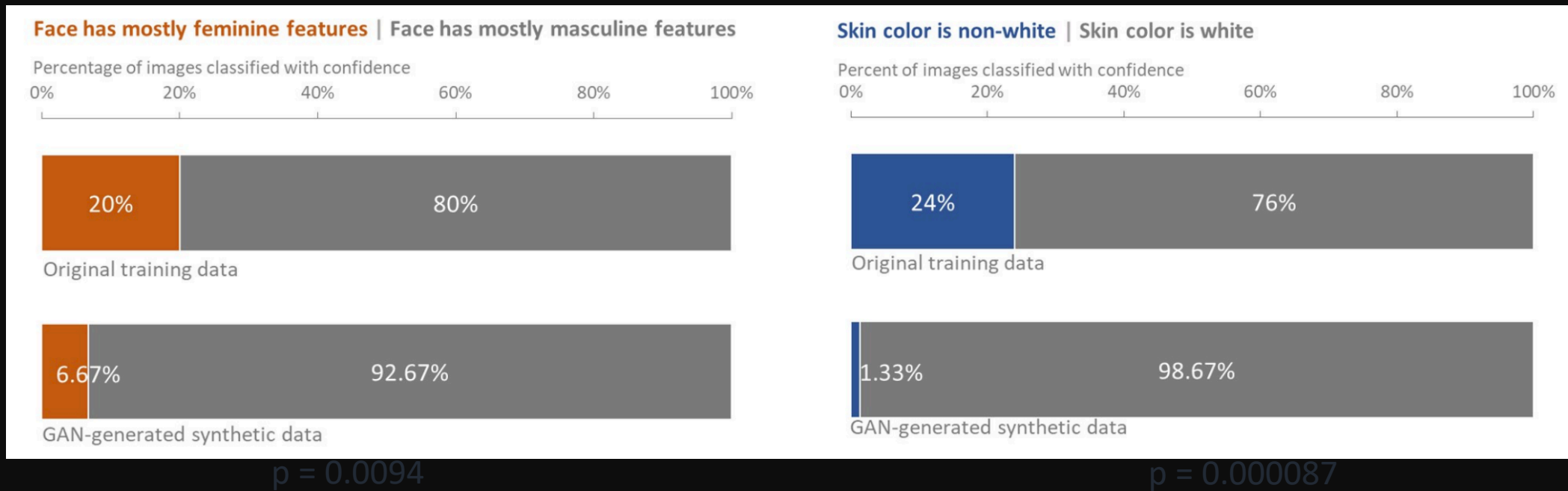
- skin color is white
- skin color is non-white
- can't tell

Between-subject design: for each distribution (p_{data} , p_{g_1} , p_{g_2} , or p_{g_3}), we ask a Turker to annotate 50 images for race and gender.

One-tailed two-proportion z-test

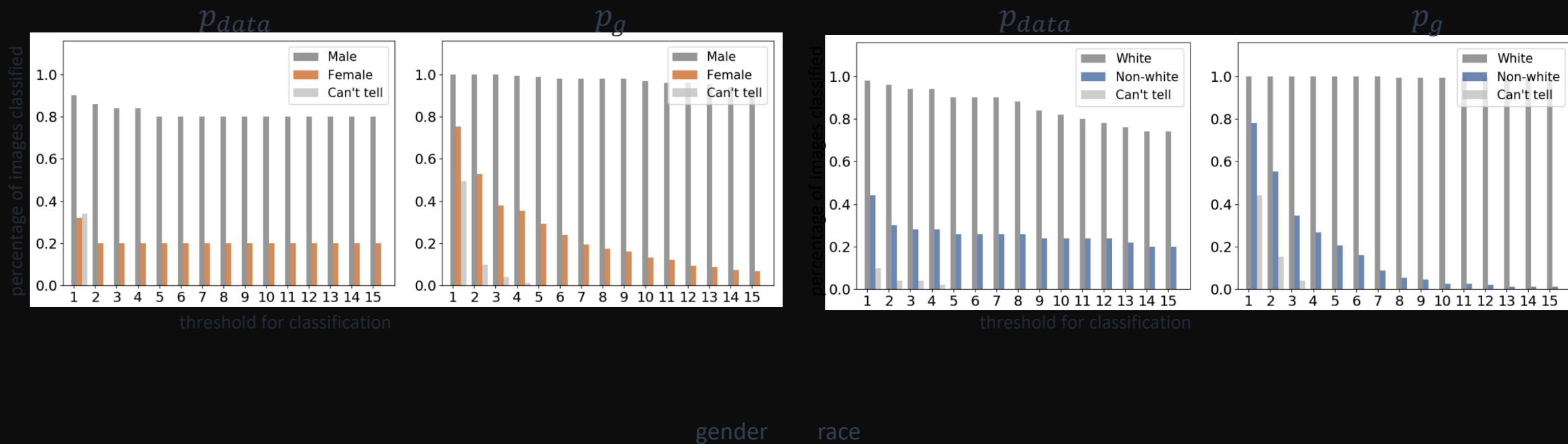
$$H_0: \hat{p} = p_0$$

$$H_a: \hat{p} < p_0$$



Using majority thresholding to label images, we find that the representation of minorities is further decreased in the synthetic data.

confidence metrics



Turkers are not as confident when generated images belong to minority classes as they are when the images belong to the majority. Is human or machine bias to blame?