The Impact of Bias in Facial Recognition Software

Towards More Ethical AI: Past, Present, and Future

By Ezra Wingard

Background ?

- In 2014, there was a big boom in conversations about AI ethics that have continued to this day[11].

- It has been demonstrated that facial recognition tech classifications on Black women and transgender people are typically much worse than on cisgender white men[2,3,4].

There are many things that are ethically alarming about using this technology!

Background ?

Real Life Ethical Dilemmas

Oliver Michael - Wrongfully arrested due to AI (2019)[6]

1. The man in the red hoodie in Michigan took a teacher's phone and broke it.
2. Said teacher shared a screenshot of video that was taken during the incident with the police.
3. AI misidentified Oliver as the man in red, and had a warrant for his arrest.
4. Oliver Michael was arrested at a traffic stop in 2019.
5. Oliver does not resemble the man physically, and was at work during the crime. He had to testify anyways.
6. Oliver said his whole life was impacted - his family, work, bills, because of this misidentification by AI.

Why does this matter?

Why should you care?

- Neural Network bias doesn’t just affect marginalized groups like transgender people - anyone can be misgendered or misclassified due to this technology.
  - There are several theories as to how and why bias/prejudices can get introduced into NNs and bias obtained from datasets is only one of them.
- Many datasets for training, validation, and testing have historically left out transgender people from such important steps.
  - This may contribute to transphobia / prejudice via NNs

Balanced vs Unbalanced datasets

What is the difference?	Balanced	Unbalanced
(Almost) equal numbers of images consisting of faces from each demographic to be measured	✓	
A lot of cisgender older white men's faces		✓
Little to no people of color's faces		✓
Mostly images of Black women's faces		✓
A couple images of trans people's faces, with mostly cisgender people's faces		✓ *

* Most commonly used datasets don't include trans people anyways.

Hypotheses:

- How accurate would a model trained on a balanced dataset (such as FairFace[1]) be on transgender individuals?
- How accurate are two different models (trained on balanced vs unbalanced datasets) on general gender classification?
- How do both unbalanced and balanced NN models perform on gender identity?

Dataset Creation Process

Instaloader API[5]

Used within Python to scrape images from Instagram (IG)

Python Script

Used to search recursively

Hashtag Input

Keywords inputted, e.g.
Transgender: #MTF
Cisgender: #Latino

Scraping

Images tagged with keywords were downloaded

Sorting

Scraped IG images were sorted into the dataset folder

Ready

Once the images were sorted, they were ready to test the NN models.

Methods ?

What did I do?

Two Testing datasets: Transgender and Cisgender faces

≈46.38%

≈53.62%

≈49.09%

≈50.9%

Total men | Total women:

192 trans men
80 unique faces

222 trans women
102 unique faces

Total men | Total women:

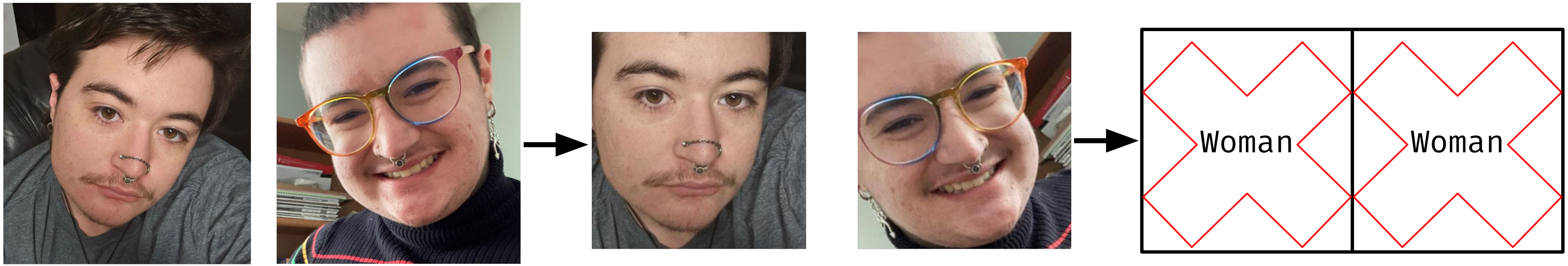
270 cis men
170 unique faces

280 cis women
162 unique faces

Binary trans people's faces were used because non-binary individuals were outside of the scope.

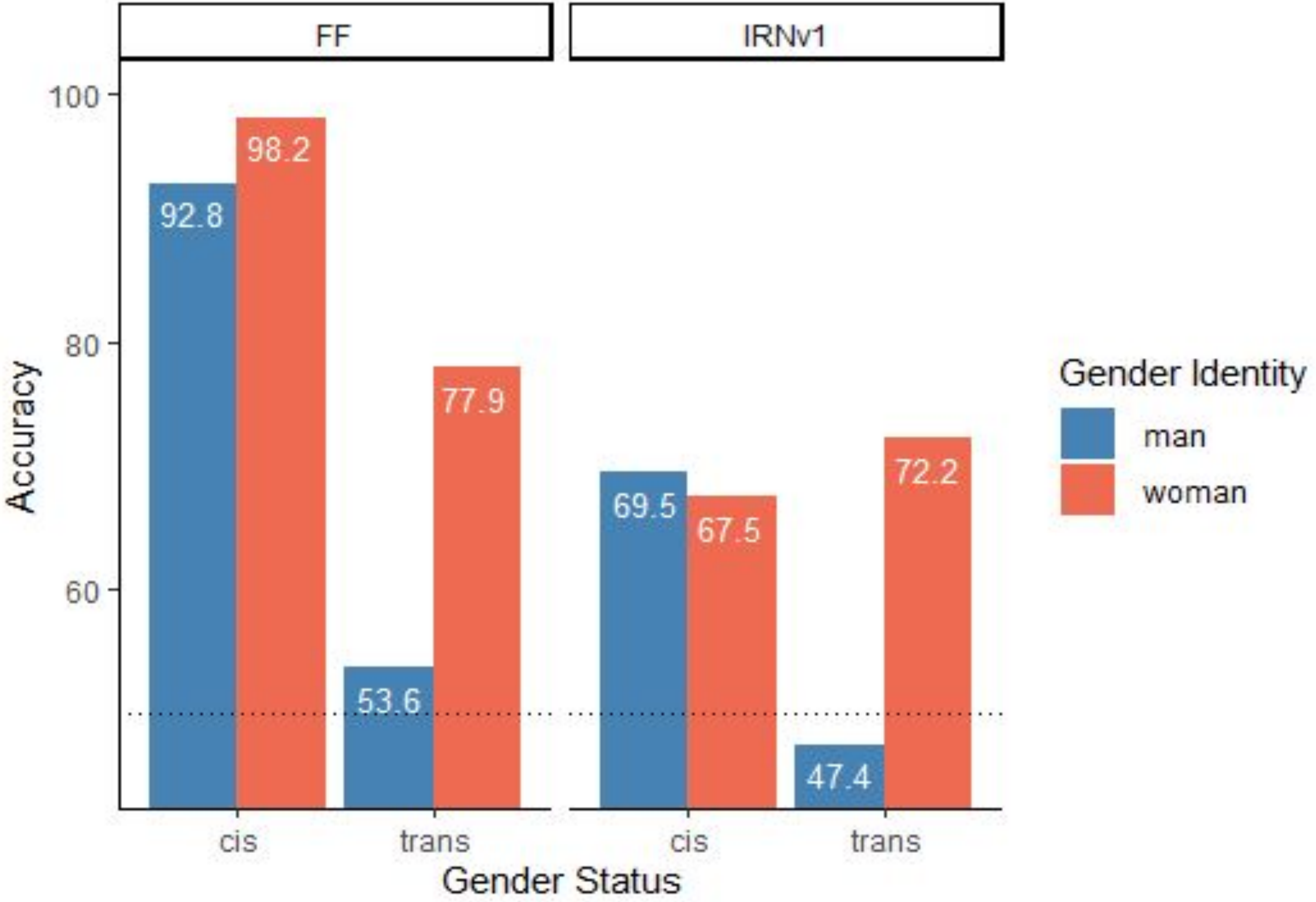
How did the NN models work?

(Using example data)



1. Initial Inputs 2. Detected Faces 3. Predictions

Results


FF
IRNv1
Accuracy
100
80
60
92.8
98.2
53.6
77.9
69.5
67.5
47.4
72.2
cis
trans
cis
trans
Gender Status
Gender Identity
man
woman

Statistics Time!

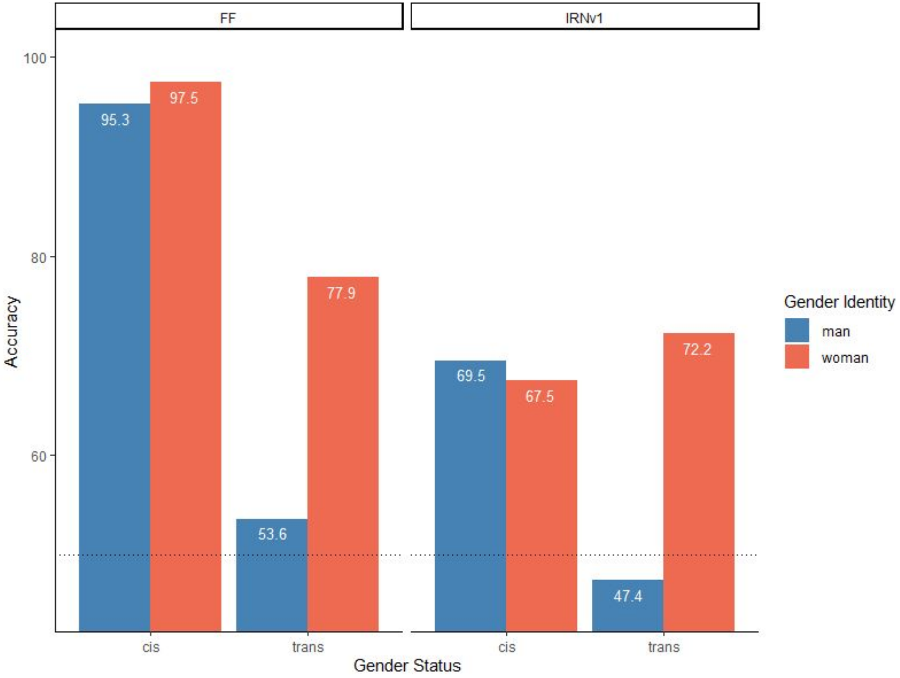
Logistic Regressions

Main Effects:

- FF 6.27x accuracy vs IRNv1.***
- Women 2.73x accuracy than men. *
- Cisgender 12.34x accuracy than trans.***

Interactions:

- FF 2.99x women accuracy than IRNv1.*
- FF 4.89x cis accuracy than trans.***
- Cis 1.12x men.~
- FF 2.83x cis men.~



~ denotes non-significant data, * significance at the p = .05 level, ** at the .001 and *** at the p < .001 level

So… What does that actually mean?

- Women of both gender statuses (transgender/cisgender) were overall more likely to be classified as women compared to men, with one exception in the IRNv1 model predictions.
- Gender status affected the accuracy rates the most in all cases (taking gender identity and model used into account), and the interaction between gender status and gender identity affected the accuracy the least. These results are extremely worrying because they demonstrate that there are tangible biases and prejudices with respect to transgender people in NN models.
- These can signal that using a balanced dataset could have helped with the accuracy rates, and using a non-balanced dataset may impede accuracy rates, especially in regards to gender status.
    - FairFace, the model trained on balanced data, did substantially better on accuracy rates overall than the unbalanced model (IRNv1).
- The effect sizes shown within the results are massive. Solely by looking at the graph, we can see that there are extreme differences between the models, gender statuses, and sometimes gender identity predictions.

What does the future hold for AI Ethics?

Gender

Masculine/Feminine spectrum

Diversity, Equity, and Inclusion

Marginalized groups must be included in more discussions

Debiasing

Keeping an eye on bias and learning how to mitigate it

Transgender vs Cisgender

Training sets to include diverse genders

Conclusion ?

"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."

— Albert Einstein

We can make AI more equitable and ethical by remembering and acknowledging the past, as well as making changes in the present to help make the future a better place for all.

Sources!

1. Kärkkäinen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).
2. Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency
3. Scheuerman, M.K., Wade, K., Lustig, C., and Brubaker, J. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. Proc. ACM Hum.-Comput. Interact. 4, CSCW1: Article 58.
4. Scheuerman, M.K., Paul, M. J., and Brubaker, J. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. Proc. ACM Hum.-Comput. Interact. 3, CSCW: Article 144
5. Instaloader. (2023). GitHub - instaloader/instaloader: Download pictures (or videos) along with their captions and other metadata from Instagram.
6. Johnson, K. (2022, March 7). *How Wrongful Arrests Based on AI Derailed 3 Men's Lives*. WIRED. https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/
7. White, M. (2014, May 1). *Wikimedia timeline of events, 2014–2016*. https://www.mollywhite.net/timelines/wikimedia/
8. davidsandberg. (2023). facenet/inception_resnet_v1.py at master · davidsandberg/facenet. https://github.com/davidsandberg/facenet/blob/master/src/models/inception_resnet_v1.py
9. Cao, Q., Shen, L., Xie, W., Parkhi, O., & Zisserman, A. (2017). VGGFace2: A dataset for recognising faces across pose and age. https://arxiv.org/abs/1710.08092#
10. Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., & Manyika, J. et al. (2022). The AI Index 2022 Annual Report. https://arxiv.org/abs/2205.03468?context=cs.AI

Questions ?

Thanks!

Are there any questions?

Resources ?

Appendix A: Logistic Regressions

* denotes significant findings at the p = .05 level, ** at the .001 and *** at the p < .001 level

Main Effects

- It was found that the gender classification accuracy was 6.27 times more likely when using the FairFace model than the IRNv1 model. ***
- Additionally, it was found that based on the gender identity, the odds of a woman being gendered correctly was 2.73 times more likely than a man. ***
- Lastly, it was found that based on the gender status, the odds of a cisgender person being gendered correctly was 12.34 times more likely than a transgender person. ***

Interactions

- FairFace was better on men than women, with the odds of a man being gendered correctly compared to the IRNv1 model being 2.99 times higher.*
- Additionally, it was found that within the FairFace, the odds of a cisgender person being gendered correctly was 4.89 times more likely than a transgender person. ***
- Lastly, the effects of cisgender status on men's accuracy was not found to be significant (1.12, p=.82), and neither was the effect of FairFace's model on cisgender men (2.83, p=0.1).

Appendix B: Model Parameters

The FairFace[1] model was used because of the appeal of it being trained on a balanced dataset.

- Pre-trained on a balanced dataset, with the following established parameters:
  - 2 genders, man and woman
  - 4 race and 7 race options
    - 4: Asian, Black, White, Indian
    - 7: White, Black, East Asian, Southeast Asian, Indian, Middle Eastern, Latine/Hispanic
  - Ages 0-100 years old, incrementing in even amounts

The InceptionResNet v1[8] model was used because of the generalizability to other NN models with unbalanced training

- Pre-trained on VGGFace2[9] - no claims of a balanced dataset, with the following established parameters:
  - 2 genders, man and woman
  - 7 race options
    - 7: White, Black, East Asian, Southeast Asian, Indian, Middle Eastern, Latine/Hispanic
  - Age was not able to be predicted from the neural network at this time.