UNIVERSITY OF SAN FRANCISCO

CHANGE THE WORLD FROM HERE

Facial Recognition for Policing

Real-life Negative Consequences of Biased Algorithms

Themes:

Algorithmic Bias and Fairness
Data Security & Privacy
Social and Economic Impact

Prerequisites:

- None for the Case Study section.
- Some knowledge of statistics, differential privacy, federated learning, and adversarial attacks for machine learning for the Technical Discussion Questions.

Owner:

Center for AI and Data Ethics at University of San Francisco

Author(s):

Hadley Dixon and Robert Clements

License:

Citation:

Dixon, Hadley and Clements, Robert. (2024). Facial Recognition for Policing: Real-life Negative Consequences of Biased Algorithms.

Objective:

The purpose of this case study is to highlight the concepts of privacy, bias and fairness in machine learning models by discussing the use of facial recognition models in policing, citing a specific example of the negative consequences of using these models without proper understanding of them and without guardrails in place.

Instructions:

1. Read through the case study individually and then answer the discussion questions as a group, or in small groups.
2. Technical audience: answer the technical discussion questions.

Case Study:

Facial recognition software offers law enforcement a powerful tool for identification and crime prevention, leveraging sophisticated deep learning models to identify individuals based on their facial features. The supposed statistical, or probabilistic, nature of machine learning generally offers government and private agencies fast-track justification, and at first glance, artificial intelligence (AI) systems seem to promise to eliminate human bias in decision-making processes. However, the use of this technology has been a topic of considerable scrutiny in recent years, particularly regarding algorithmic bias and fairness. As both theoretical and real-world examples show, human bias can be perpetuated and even exacerbated by AI. Similarly, its use has raised concerns regarding civil liberties and equity in the criminal legal system (Najibi, 2020).

In 2019, a Black New Jersey resident was misidentified by facial recognition technology and falsely jailed. Nijeer Parks filed a lawsuit against his prosecutors for his wrongful arrest and imprisonment for hotel theft, after inaccurate identification by facial recognition software. In January 2019, Parks was accused of shoplifting from a Hampton Inn location and then subsequently evading the police. Parks denied all claims, arguing that he did not own a car at the time and never possessed a driver's license. Moreover, Parks claims to have never been in Woodbridge, the city where the crime occurred.

Both evidence and forensics back up Parks' claims, as well as a solid alibi which suggested Parks had nothing to do with the crime. Despite this, Parks sat in jail for 10 days. Prosecutors refused to check fingerprints and DNA at the scene of the crime, instead relying solely on the results of facial recognition software, which was illegally used during prosecution. Officers submitted a blurry photo from a fake driver's license, which was presented at the Hampton Inn by the perpetrator before fleeing the hotel. The results from the software returned Parks' arrest photo from several years prior as a "possible hit" (ACLU, 2024). The prosecution was wholly satisfied with the results returned by the software, and circumvented all standard investigative procedures. Soon after, the ACLU voiced their support of Parks, stating how his case "represents the unfortunate and increasingly common story of how the police's uncritical reliance on

results of unreliable FRT searches can deprive the innocent of their liberty and directly violate constitutional rights" (Difillipo, 2024).

Across the country, misidentification in facial recognition technology disproportionately impacts people of color and minority communities. NIST (2019) finds that "Asian and African American individuals were up to 100 times more likely to be misidentified compared to White males. Women were more likely to be misidentified than men (Buolamwini, 2018). Middle-aged White males had the highest accuracy rates across various facial recognition algorithms." The better performance of facial recognition on white male faces is likely linked to overrepresentation of white male faces in the dataset used for training the model. In the case of Parks, the model found a match between a blurry photo and his photo in a database of mugshots, a database that likely has an overrepresentation of Black male faces, reflecting historical racial biases within America's criminal justice system. The combined use of a model that might perform worse on recognizing Black faces and a matching database containing a relatively high proportion of Black faces, along with an over-reliance on the accuracy of these models and general lack of understanding of the existence of algorithmic bias, is what led to the arrest of Parks.

Parks' case is not an outlier; bias in facial recognition technologies is common. Google broke news in 2015 for mistakenly tagging Black people as "Gorillas" in the search feature of their Google Photos app, shining a harsh light on the shortcomings of facial recognition (Barr, 2015). These examples, along with many others, showcase the dangers of blindly trusting facial recognition, especially when used for policing. Similarly, current AI facial recognition allows for more precise discrimination in areas already subject to heavy policing. Marginalized communities, including people of color, immigrants, and low-income individuals, are often subject to higher levels of police surveillance and are more likely to be falsely identified and unfairly arrested (Kutateladze, 2014). Concurrently, police departments are using facial recognition more and more often and for less serious crimes. These systems can and will sometimes return false positives, leading to wrongful arrests and possibly convictions. When facial recognition technology is deployed in these areas with existing discriminatory policing practices, it can reinforce these biases, pointing to the crucial need to reevaluate and mitigate harms to ensure fair treatment.

Discussion Questions:

1. Are facial recognition models biased, discriminatory and/or unfair?
2. What level of accuracy would facial detection models need to achieve for you to feel comfortable relying solely, or heavily, on them as evidence to implicate someone in a crime?
3. What strategies, if any, can be implemented by law enforcement to minimize the effects of both human and algorithmic bias?
4. What are some possible solutions to reducing bias in facial recognition technologies? How might bias and fairness be assessed?

Facial recognition technology would not be possible without access to a large database of high quality photos of faces. There are many datasets of faces available that are commonly used in facial recognition and generative modeling research and model training, such as the Flickr-Faces-HQ (FFHQ) (Karras, 2019), VGGFace2 (Cao, 2018), and many others. Most of these that are publicly available have open non-commercial licenses, meaning that they are not meant to be used for commercial purposes. While FFHQ claims that "when collecting the data, we were careful to only include photos that – to the best of our knowledge – were intended for free use and redistribution by their respective authors" they offer a way to opt out, recognizing that some people may be unaware that their face is included in this dataset. Most other publicly available face datasets do not make such claims, they are simply scraped from the "open" web, and are therefore already public and legal, according to the dataset creators.

There are also commercially available facial recognition products, one of the more controversial being offered by the company Clearview AI. Clearview has attained a massive database of face images scraped from the web (approximately 40 billion) including from social media and mugshot websites. Clearview has been forced to pay several fines across European countries for being in violation of data privacy laws, and has been banned in the United Kingdom (Heikkilä, 2022). In the United States, a country without a national data privacy law, Clearview has been banned from selling its products to private companies as the result of a lawsuit brought forth by the ACLU with respect to Clearview's violation of data privacy laws in the state of Illinois, but is not restricted from selling to law enforcement, with the exception of a few cities and the state of Illinois (Clayton, 2023).

Discussion Questions:

1. Should the United States adopt a national data privacy law, and if so, to what extent should the use of facial recognition technology be regulated? Should it be banned, or should there be other less strict restrictions put in place?
2. Under what circumstances, if any, should law enforcement have access to a comprehensive database of face images? How can we balance the benefits of public safety with personal privacy?
3. How can we protect the privacy rights of more vulnerable groups, such as children, elderly, unhoused, refugees, or persons with disabilities?

Technical Discussion Questions:

1. What are the potential sources of unfairness in facial recognition models? What solutions exist for collecting a more representative sample of training data or otherwise correct for fairness?
2. There are privacy-preserving methods for training ML models, such as differential privacy or federated learning, that support data security. Differential privacy

anonymizes the raw data, while federated learning trains models without having to store all data together in one place, meaning the data can stay on a specific device such as a smartphone without being shared with other devices. Are these approaches enough to mitigate the privacy concerns of facial recognition models?

3. Facial recognition models are vulnerable to adversarial attacks, i.e. attacks that trick the model into making false predictions. What are some potential examples of adversarial attacks, their implications, and how might they be remedied?

References:

Barr, A. (2015, July 1). *Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms*. The Wall Street Journal. https://www.wsj.com/articles/BL-DGB-42522

Buolamwini, J. and Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1–15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., Zisserman A. (2018). VGGFace2: A dataset for recognising faces across pose and age. International Conference on Automatic Face and Gesture Recognition, 2018. https://www.robots.ox.ac.uk/~vgg/publications/2018/Cao18/

Carina, W. (2022). Failing at Face Value: The Effect of Biased Facial Recognition Technology on Racial Discrimination in Criminal Justice. Scientific and Social Research. (full article here: https://drive.google.com/drive/folders/1i6CFh-ls-10AYcGIP4mPIJhH-SyCcvMh)

Clayton, J. and Derico, B. (2023) Clearview AI used nearly 1m times by US police, it tells the BBC. BBC News, San Francisco. https://www.bbc.com/news/technology-65057011

Difilippo, D. (2024, February 1). *Lawsuit seen as crucial test of police use of facial recognition technology*. New Jersey Monitor. https://newjerseymonitor.com/2024/02/01/lawsuit-seen-as-crucial-test-of-police-use-of-facial-recognition-technology/

Heikkilä, M. (2022) The walls are closing in on Clearview AI
The controversial face recognition company was just fined $10 million for scraping UK faces from the web. That might not be the end of it. MIT Technology Review. https://www.technologyreview.com/2022/05/24/1052653/clearview-ai-data-privacy-uk/

Karras, T., Laine, S., and Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. https://arxiv.org/pdf/1812.04948.pdf

Kutateladze, B. L., Andiloro, N.R., Johnson, B, D., & Spohn, C. C., (2014, August). Cumulative Disadvantage: Examining Racial and Ethnic Disparity in Prosecution and Sentencing. *Criminology*, *52*(3), 514–551. https://doi.org/10.1111/1745-9125.12047.

Li, D. (2020, December 29). *Black man in New Jersey misidentified by facial recognition tech and falsely jailed, lawsuit claims.* NBC News. https://www.nbcnews.com/news/us-news/black-man-new-jersey-misidentified-facial-recognition-tech-falsely-jailed-n1252489

Najibi, A. Harvard Graduate School of Arts and Sciences. (2020, October 24). *Racial Discrimination in Face Recognition Technology*. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

National Institute of Standards and Technology. (2019). NIST study evaluates effects of race, age, sex on face recognition software. Retrieved August 29, 2022, from https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software