Chapter 10
Fairness in ML

Slides (pdf) Video (Panopto)

In this chapter, we will look at more examples of bias and fairness in Machine Learning. We will dive deeper into the motivating examples for why one should care about bias, multiple case studies of bias sources, the pillars of trustworthy machine learning, how to tackle bias in machine learning, and finally real world examples of tackling bias.

10.1 Examples of Bias

10.1.1 COMPAS Recidivism

COMPAS is a tool used to predict **recidivism**, the tendency of a convicted criminal to reoffend and it has been used in almost all 50 states. In 2016, ProPublica investigated the bias COMPAS exhibited against black people and concluded a statistically signficant difference in prediction scores for people of different races committing the same crimes.

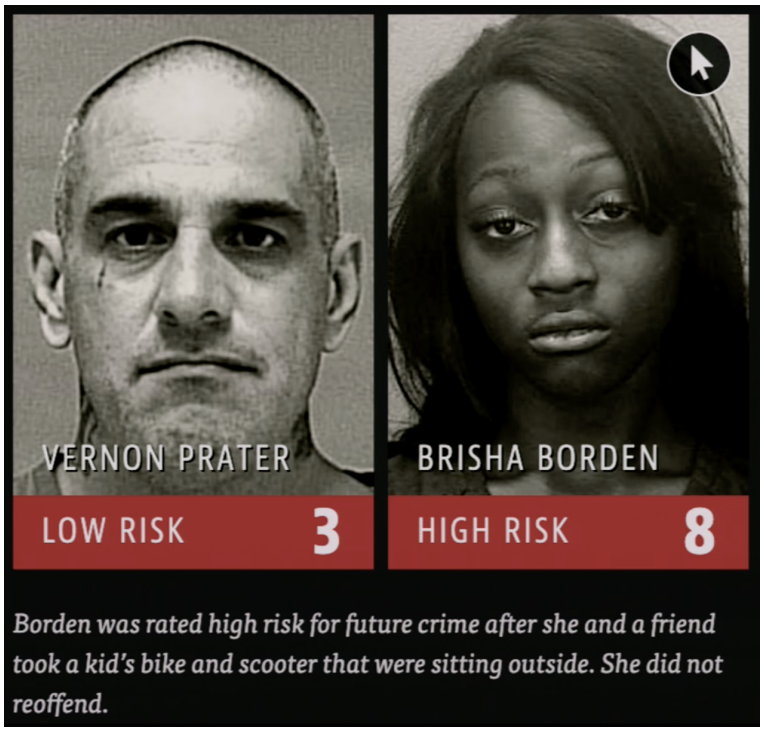

Figure 10.1: Shown above is one example out of the thousands investigated by ProPublica of criminals of the same offense assigned different recidivism scores. Both Prater and Borden were convicted of stealing a bike. While Prater was given a recidivism score of "3", Borden received an "8".

10.1.2 Amazon Recruiting Algorithm

Amazon discovered in 2015 that their recruiting algorithm was biased against women. Their model was trained on the résumés of previously hired employees, so whether a new candidate passed the primary screening was dependent on how the model rated their résumé. The company couldn't find a way to fix the bias in their model so they got rid of its use completely.

10.1.3 Healthcare Risk Management

In 2019 the Scientific American published an article on an unnamed software deployed by American healthcare companies to determine which patients qualified for "high-risk care management". The software was used on 200 million patients in the US. Frequency of health care usage and medical spending were factors the model used to determine qualifying patients, and as a result black patients, while suffering from chronic illness at higher rates, would qualify less.

10.1.4 Predictive Policing

PredPol, a software company specialized in predicting policing, made predictions on future crime based on existing crime data. Criticized by several academics, PredPol was found to be target black and latinx communities. As more police were deployed in certain neighborhoods, members of those neighborhoods were watched more frequently and thus arrested for more crimes, even if these crimes were minor. This data was used to further train the model, creating a feedback loop, thereby further biasing the model.

10.1.5 Amazon Prime

Amazon Prime's one day delivery service was found to exclude predominantly black neighborhoods across the United States. Neighborhood income was used to determine where Amazon warehouses should be built. As Amazon is a large company, it quickly came under fire by major news outlets. To avoid a bad public image, the company quickly vowed to fill racial gaps in its delivery services.

10.1.6 Facebook Ads

In 2019 MIT published a study showcasing Facebook's biased algorithm for ads. While men were more likely to be shown ads for careers like being a doctor or an engineer, women were shown ads for nursing and becoming secretaries. While white people were shown ads for home sales minorities were shown rentals.

10.1.7 Google Ads

In 2013 Latanya Sweeny, a Harvard Professor, found out that Googling "black-sounding" names were 25% more likely to trigger ads for finding criminal records in someone's background.

10.1.8 Face Detection

Computer scientist and activist Joy Buolamwini began a project that became her MIT Thesis: *Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers.*

Figure 10.2: Harvard Professor Dr. Latanya Sweeney.

The research found that many of the leading facial classification technologies used in a variety of applications today classify black women's faces nearly 30% less accurately than their white male counterparts.

Figure 10.3: Activist and computer scientist Joy Buolamwini. You can watch Buolwamwini's interview here.

10.1.9 Nikon Camera Blink Detection

Nikon's blink detection was found to falsely classify Asian people as blinking.

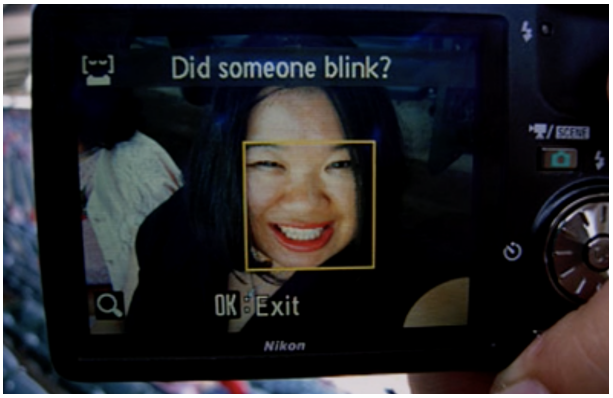

Figure 10.4: Nikon's failure to correctly detect Asian faces.

10.2 Source of Bias: Case Studies

Now that we have gone over dire examples of why fairness in ML matters, we will examine some of the **sources** of bias. First, let's review some different causes that can lead to bias. Recall that there is **historical bias**, **measurement bias representation bias**, **deployment bias**, **evaluation bias**.

We will look at examples of case studies where such biases are found.

Example(s)

Case Study: Distaster Relief
Goal: Accurately assess damage and send appropriate relief resources.
Data: Twitter posts, facebook posts goecoded with coordinates within disaster area, keywords and hashtags related to the storm.
Analysis: Intensity and type of damage by neighborhood.
Actions: Assessment and allocate relief effort (type and amount).
Constraints: Limited resources for relief efforts.

Can you find some problems with this method of resource allocation?

1. **Representation bias:** Not everyone has access to social media.
2. **Historical bias:** Wealthier neighborhoods, with more infrastructure, will report more damage.

Example(s)

Case Study: Personality Detection
This case study is based on a real example called Faceception, an Israeli startup is selling a software that can classify personalities with 80% accuracy.
Goal: Detect if someone is a terrorist or not.
Data: People's faces.
Analysis: IQ, profession, skills, personality, terrorist activity.
Actions: US Homeland Security body purchased the software in 2016.
Constraints: Limited pool of faces and personality for training data.

Can you find some problems with this method?

1. **Representation bias:** One country is not representative of the faces and personalities around the world.
2. **Historical bias:** In media headlines middle eastern people are labeled as "terrorists" while white terrorists are always labeled as "troubled young man with mental health problems". If this is the labeling used for the dataset, then the model is trained on skewed historical bias.

Example(s)

Case Study: Sexuality Detection
This case study is also based on a real example. A Stanford professor developed a facial recognition system using 75000 photos from online dating profiles of all white people to label people's sexualities. The model reaches 71% accuracy for women and 81% accuracy for men.

Goal: Detect if someone is gay or straight.
Data: People's faces.
Analysis: Sexuality based on facial structure.
Actions: The professor claims that the research would help people, but he said in an interview that he was "flattered" that the Russian Prime Minister and the Russian cabinet convened with him to discuss his technology. In 2018 Christopher Wylie, a whistleblower, revealed that Cambridge Analytica was using this technology as an "instrument of psychological warfare".

The problems and constraints with this application of ML are obvious.

10.3 Pillars of Trustworthy ML

As you can see, machine learning is a powerful double-edged sword. While the technology has the potential to do good, it can also have disastrous consequences if handled by the wrong people. As an ML developer in the making, it is your responsibility to employ the Pillars of Trustworthy ML so that we can transform this space into one that honors an intersectionally empowering future.

10.3.1 Pillar 1: Privacy

Cloud photo storage company Ever AI sold the contents of thousands of people's private camera rolls to surveillance agencies to train facial recognition software. Upon accusations, Ever AI renamed itself to Paravision to claim that they had never done what Ever AI did. Before 2020 they made $29 million in profits before finally busted by the ACLU. Thus, privacy is imperative for an ethical ML system.

10.3.2 Pillar 2: Ethics

In 2018, CEO Matthew Zeiler disclosed that computer vision software company Clarifai was selling software to the Pentagon for autonomous weapons. Clarifai was aiming to use computer vision technologies to auto-detect enemies on a battlefield. Zeiler was forced to disclose the news thanks to company whistleblower Liz O'Sullivan. By whistleblowing the usages of Clarifai, O'Sullivan exemplified ethics in ML.

10.3.3 Pillar 3: Interpretability

Interpretability is the ability to explain *why* a model makes a certain decision. An example of an interpretability problem is when a bank uses an ML model to decide which customers' loan request should be approved. A customer who earnestly needed money asks why their loan request did not get approved and what they can do to get it approved. To answer to this customer, the bank would have to dissect and understand how its model made its decision. This pillar is especially relevant to the healthcare management example we previously showed.

10.3.4 Pillar 4: Robustness

Robustness will test how consistently reliable a model is. While an image of a pig is classified correctly at first, adding random noise to the image confuses the model and the model classifies the pig as an airplane. This is an especially large problem for facial detection if a model is not exposed to a diverse enough set of face types.

10.4 Tackling Bias

"Fairness" can take on a number of definitions depending on the context. In order to conduct a context-aware bias mitigation, we can follow the practices of each stages of processing:

1. Pre-processing: Resampling of rows of the data, reweighing rows of the data, flipping the class labels across groups, and omitting sensitive variables or proxies.
2. : In-processing: Modifying the loss function to account for fairness constraints with respect to an analysis of false positives and false negatives.
3. Post-processing: Adjust the outputs of the model to abide by a fairness criteria.

10.5 Real World Examples of Tackling Bias