Introduction

Premises

- Others' decisions can profoundly affect the course of our lives, such as decisions to admit us to a particular school, offer us a job, or grant us a mortgage (Barocas et al., 2023, p.1).
- Arbitrary, inconsistent, or faulty decision-making raises serious concerns (Barocas et al., 2023, p.1).
- Data-driven decisions are often more accurate than decisions based on intuition or expertise in head-to-head comparisons on fixed tasks. A body of research has compared the accuracy of statistical models to human judgments, including those of experts with years of experience (Barocas et al., 2023, p.1).
- ML promises to bring greater discipline to decision-making by uncovering factors relevant to decision-making that humans might overlook due to complexity or subtlety. ML allows us to defer the question of relevance to the data themselves, determining which observed factors bear a statistical relationship to the outcome (Barocas et al., 2023, p.1).
- ML is increasingly used in high-stakes decision-making (e.g., criminal justice, employment, lending).
- The ML pipeline consists of several stages: measurement, learning, prediction, action, and feedback.
  - Measurement: The process of converting the state of the world into data, involving subjective human decisions. Measurement involves defining variables, collecting data, and understanding the data's origin and context.
  - Learning: The stage where data is used to create a model that generalizes patterns from the training data.
  - Action: The use of the model's predictions to make decisions about new, unseen inputs. Predictions can involve classification (e.g., spam detection), regression (e.g., risk scoring), or information retrieval (e.g., search results).
  - Feedback: Machine learning systems may use user reactions to refine models.
- Reliable learning requires good data: a large number, diverse appearances, and well-annotated sets (Barocas et al., 2023, p.2).

Proposition ML being "evidence-based" does not guarantee accurate, reliable, or fair decisions. Even well-intentioned applications of machine learning might give rise to objectionable results perpetuating biases and injustices (Barocas et al., 2023, p. 3).

Reasons

- Modeling human behavior with machine learning often reflects historical prejudices, stereotypes, and inequalities (Barocas et al., 2023, p.2). Training data reflects real-world disparities, distortions, and biases.
  - A 2016 investigation found demographic disparities in Amazon's qualifying neighborhoods (which neighborhoods receive free same-day delivery), with White residents more likely than Black residents to be included (Barocas et al., 2023, p.3). Amazon argues its system is justified by efficiency and cost, not race, but it still results in racially disparate opportunities.
  - Researchers sometimes repurpose existing classification schemes (e.g., ImageNet) that may contain outdated associations, such as occupations that no longer exist and stereotyped gender associations (Barocas et al., 2023, p.9).
  - Camera technology historically biased toward lighter skin tones due to design choices.
  - Translating from English to Turkish and back introduces gender stereotypes due to biases in training data.

- Human decision-makers do not always maximize predictive accuracy; they may consider morally relevant factors (Barocas et al., 2023, p.2).
  - Judges may avoid using certain statistical factors (e.g., age) in sentencing, viewing younger defendants as less culpable (Barocas et al., 2023, p.2).
  - More generally, ML doesn't know good or bad: Some patterns in training data represent valuable knowledge, while others reflect harmful stereotypes. Learning algorithms cannot inherently distinguish between the two, as both are shaped by social norms. Without specific intervention, machine learning will extract stereotypes just as it extracts knowledge (Barocas et al., 2023, p.10).
- Unlike humans, automated decision-making may produce absurd outcomes due to errors in data (Barocas et al., 2023, p.3).
  - Evidence?
- ML practitioners frequently need to define new target variables (e.g., recidivism, job performance). Bias in defining target variables is critical because it can distort predictions relative to the intended construct (Barocas et al., 2023, p.8).
  - Target variables like "creditworthiness" are constructs, not intrinsic properties, and are difficult to measure objectively.
  - Performance review scores as a measure of a "good employee" inherit biases from evaluators.
  - Using proxies (e.g., arrests as a measure of crime) introduces distortions from biased policing.
- Moving from individual data points to datasets introduces an additional layer of potential biases.
  - Image datasets are used to train computer vision systems for tasks like object recognition. If these datasets were representative of the underlying visual world, a model trained on one dataset should perform well on another. In practice, there is a significant drop in accuracy when models are trained on one dataset and tested on another (Barocas et al., 2023, p.9).
- No easy fix: Simply withholding, e.g., gender from data does not eliminate gender bias because of proxies and redundant encodings.
  - Attributes like age at which someone starts programming can serve as proxies for gender and perpetuate bias. How long someone has been programming is a factor that gives us valuable information about their suitability for a programming job, but it also reflects the reality of gender stereotyping (Barocas et al., 2023, p.11).
- ML models perform worse for minority groups due to sample size disparities. Generalizing based on the majority culture leads to high error rates for minority groups (Barocas et al., 2023, p.12).
  - Anomaly detection in ML shows high error rates for minority groups (e.g., detecting uncommon names as fake).