(Informal) Logic: Barocas et al. Ch. 1

WRIT 0590: Module 2.2

Nikita Bezrukov
University of Pennsylvania
nikitab@sas.upenn.edu

January 21, 2025

Roadmap

Why Fairness Matters

- **Accuracy vs. fairness:** Data-driven decisions often outperform human intuition on specific tasks, yet we worry about potential biases.
- **High-stakes decisions:** Admissions, hiring, lending, and sentencing profoundly affect people's lives.
- **Potential harm:** Faulty or biased decision-making—whether human or algorithmic—can perpetuate injustice.

ML in Decision-Making

- **Promise of ML:**
  - Uncover subtle factors humans might miss
  - Potentially more "objective" or evidence-based
- **Reality of Bias:**
  - ML learns from historical data, which can carry forward existing inequities
  - Inherent complexities around measuring human constructs (e.g., "creditworthiness")

Steps in the ML Pipeline

1. **Measurement:** Converting the real world into data.
   - Subjective choices: which variables to collect
   - Potential distortion or bias in data gathering
2. **Learning:** Using data to train a model.
   - Patterns are extracted—good or bad
   - Reflects statistical relationships in training data
3. **Prediction & Action:**
   - Classification, regression, or ranking tasks
   - Used for real-world decisions (e.g., lending, hiring)
4. **Feedback:**
   - User outcomes or responses refine the model
   - Risks reinforcing existing patterns if feedback is also biased

Key Concern

Even well-intentioned ML applications can yield objectionable outcomes by perpetuating biases and injustices.

- Historical biases in data
- Inadequate or skewed measurement
- Complex moral values overlooked

Examples of Algorithmic Bias

- **Geographic disparity:** Amazon's free same-day delivery once excluded predominantly Black neighborhoods
  - The company cited efficiency and cost, but the impact was racially skewed.
- **Repurposed classification schemes:** Datasets like ImageNet may contain outdated or stereotyped labels (e.g., gendered roles).
- **Language translation:** Translating between certain languages introduces gender stereotypes, reflecting biases in training text.

Moral and Social Factors

- **ML can't distinguish:**
  - It will learn both helpful patterns and harmful stereotypes unless we intervene.
- **Human judgment vs. ML:**
  - Judges may refuse to consider "predictive" factors like age in sentencing because of moral considerations.

Target Variables

- **Constructs vs. reality:**
  - “Creditworthiness” and “job performance” are human-defined concepts
  - Often rely on proxies (e.g., arrests for crime)
- **Risk of bias:**
  - Biased policing leads to over-representation of certain groups in arrest data
  - Performance reviews might reflect supervisor stereotypes

Dataset-Level Challenges

- **Shifts in distributions:**
  - A model trained on Dataset A often fails on Dataset B.
- **Sample size disparities:**
  - Minority groups are underrepresented, leading to higher error rates.
- **Proxies and redundant encodings:**
  - Withholding “gender” does not remove bias—features like “age at first coding experience” may inadvertently reveal it.

No Easy Fix

- **Hiding sensitive attributes**
  - Insufficient due to proxies in the data
- **Improving data diversity**
  - Helps, but doesn't magically remove historical biases
- **Awareness of moral judgment**
  - Some patterns are ethically off-limits or context-dependent

Fairness requires ongoing social, technical, and legal efforts.

Conclusion

- ML can amplify historical inequalities if not designed and monitored carefully.
- Ensuring fairness means critically examining:
  - How data is collected and measured
  - What target variables and proxies are used
  - Which moral and social factors are (or should be) ignored or included
- **Key message:** Data-driven does not automatically mean objective or fair.