Prediction-based decisions & fairness: choices, assumptions, and definitions

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum

November 12, 2019

Prediction-based decisions

- Industry
  - lending
  - hiring
  - online advertising
- Government
  - pretrial detention
  - child maltreatment screening
  - predicting lead poisoning
  - welfare eligibility

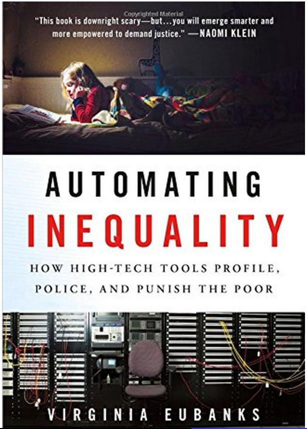
Things to talk about

- Choices to justify a prediction-based decision system
- 4 flavors of fairness definitions
- Confusing terminology
- "Conclusion"

Choices to justify a prediction-based decision system

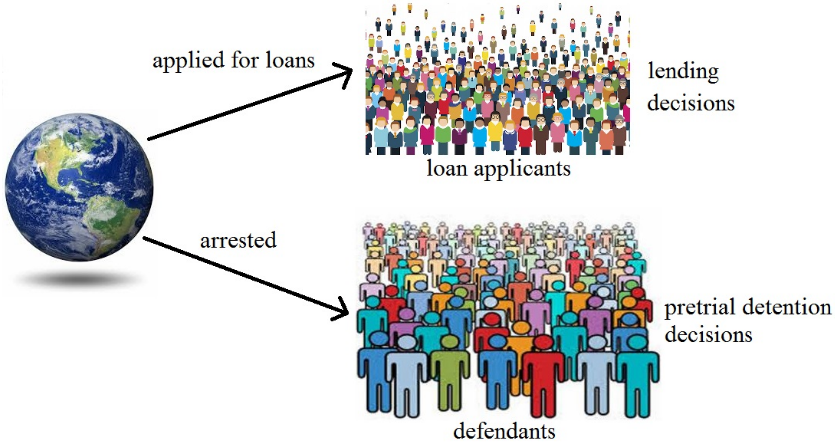
1. Choose a goal

- Company: profits
- Benevolent social planner: justice, welfare
- Often goals conflict (Eubanks, 2018)
- Assume progress is summarized by a number ("utility"): G



2. Choose a population

- Who are you making decisions about?
- Is the mechanism of entry into this population unjust?



3. Choose a decision space

- Assume decisions are made at the individual level and are binary
  - d_i = lend or not
  - d_i = detain or not

3. Choose a decision space

- Assume decisions are made at the individual level and are binary
  - d_i = lend or not
  - d_i = detain or not
- Less harmful interventions are often left out
  - longer-term, lower-interest loans
  - transportation to court, job opportunities

Advance Peace

Advance Peace is dedicated to ending cyclical and retaliatory gun violence in American urban neighborhoods. We invest in the ***development***, ***health***, and ***wellbeing*** of those at the center of this crisis.

4. Choose an outcome relevant to the decision

- d_i = family intervention program or not
- y_i = child maltreatment or not

4. Choose an outcome relevant to the decision

- d_i = family intervention program or not
- y_i = child maltreatment or not
  - Family 1: maltreatment with or without the program
  - Family 2: maltreatment without the program, but the program helps

4. Choose an outcome relevant to the decision

- d_i = family intervention program or not
- y_i = child maltreatment or not
    - Family 1: maltreatment with or without the program
    - Family 2: maltreatment without the program, but the program helps
    - **Enroll Family 2 in the program, but Family 1 may need an alternative**
    - $\Rightarrow$ **consider both** *potential outcomes*: $y_i(0), y_i(1)$

4. Choose an outcome relevant to the decision

- Let $y_i(\mathbf{d})$ be the potential outcome under the whole decision system
- Assume utility is a function of these and *no other outcomes*: $G(\mathbf{d}) = \gamma(\mathbf{d}, \mathbf{y}(\mathbf{0}), ..., \mathbf{y}(\mathbf{1}))$
- e.g. Kleinberg et al. (2018) evaluate admissions in terms of future GPA, ignoring other outcomes

5. Assume decisions can be evaluated separately, symmetrically, and simultaneously

- Separately
  - No interference: $y_i(\mathbf{d}) = y_i(d_i)$
  - No consideration of group aggregates

5. Assume decisions can be evaluated separately, symmetrically, and simultaneously

- Separately
- Symmetrically
  - Identically
  - Harm of denying a loan to someone who can repay is equal across people

5. Assume decisions can be evaluated separately, symmetrically, and simultaneously

- Separately
- Symmetrically
- Simultaneously
    - Dynamics don't matter
      (Harcourt, 2008; Hu and Chen, 2018; Hu et al., 2018; Milli et al., 2018)

5. Assume decisions can be evaluated separately, symmetrically, and simultaneously

- Separately
- Symmetrically
- Simultaneously

$\Rightarrow$

$$
\begin{aligned}
G^{\text{sss}}(\mathbf{d}) &\equiv \frac{1}{n}\sum_{i=1}^{n} \gamma^{\text{sss}}(d_i, y_i(0), y_i(1)) \\
&= E[\gamma^{\text{sss}}(D, Y(0), Y(1))]
\end{aligned}
$$

6. Assume away one potential outcome

- Predict crime if released: $y_i(0)$
  Assume no crime if detained: $y_i(1) = 0$

- Predict child abuse without intervention: $y_i(0)$
  Assume intervention helps: $y_i(1) = 0$

- But neither is obvious


Can an Algorithm Tell When Kids Are in Danger?

Child protective agencies are haunted when they fail to save kids. Pittsburgh officials believe a new data analysis program is helping them make better judgment calls.

By DAN HURLEY JAN. 2, 2018

7. Choose the prediction setup

- Let Y be the potential outcome to predict

$$
\begin{aligned}
G^{sss}(\mathbf{d}) &= E[\gamma^{sss}(D, Y)] \\
&= E[g_{TP} Y D \qquad + g_{FP}(1 - Y)D \\
&\quad + g_{FN} Y(1 - D) \qquad + g_{TN}(1 - Y)(1 - D)]
\end{aligned}
$$

7. Choose the prediction setup

- Rearrange, drop terms without D:

$$G^{\text{sss},*}(\mathbf{d};\mathbf{c}) \equiv E\left[YD - \underbrace{\frac{g_{\text{TN}} - g_{\text{FP}}}{g_{\text{TP}} + g_{\text{TN}} - g_{\text{FP}} - g_{\text{FN}}}}_{\equiv \mathbf{c}} D\right]$$

- maximizing $G^{\text{sss},*}(\mathbf{d}; 0.5) \Leftrightarrow$ maximizing accuracy $P[Y = D]$

7. Choose the prediction setup

- Decisions must be functions of variables at decision time: $D = \delta(V)$
- $G^{sss,*}(\delta; c) = E[Y\delta(V) - c\delta(V)]$ is maximized at

$$\delta(v) = I(P[Y = 1|V = v] \geqslant c)$$

- *single-threshold rule*

7. Choose the prediction setup

- *Variable selection*: $\mathrm{P}[Y = 1 | V = v]$ changes with choice of V

7. Choose the prediction setup

- *Variable selection*: $P[Y = 1|V = v]$ changes with choice of V
- *Sampling*:
  - sample to estimate $P[Y = 1|V = v]$
  - non-representative sample can lead to bias

7. Choose the prediction setup

- *Variable selection*: $\mathrm{P}[Y = 1 | V = v]$ changes with choice of V
- *Sampling*:
  - sample to estimate $\mathrm{P}[Y = 1 | V = v]$
  - non-representative sample can lead to bias
- *Measurement*: e.g. Y is defined as crime, but measured as arrests

7. Choose the prediction setup

- *Variable selection*: $P[Y = 1|V = v]$ changes with choice of V
- *Sampling*:
  - sample to estimate $P[Y = 1|V = v]$
  - non-representative sample can lead to bias
- *Measurement*: e.g. Y is defined as crime, but measured as arrests
- *Model selection*: estimate of $P[Y = 1|V = v]$ changes with choice of model

What about fairness?

- Consider an advantaged ($A = a$) and disadvantaged ($A = a'$) group

What about fairness?

- Consider an advantaged ($A = a$) and disadvantaged ($A = a'$) group
- Under *many* assumptions, single-threshold rule maximizes utility per group. Fair?
  - Disadvantaged group could have a lower maximum
  - Impacts of decisions may not be contained within groups

What about fairness?

- Consider an advantaged ($A = a$) and disadvantaged ($A = a'$) group
- Under *many* assumptions, single-threshold rule maximizes utility per group. Fair?
  - Disadvantaged group could have a lower maximum
  - Impacts of decisions may not be contained within groups
- People with the same estimates of $P[Y = 1|V = v]$ are treated the same. Fair?
  - Conditional probabilities change with variable selection
  - Estimates depend on sample, measurement, models

What about fairness?

- Consider an advantaged ($A = a$) and disadvantaged ($A = a'$) group
- Under *many* assumptions, single-threshold rule maximizes utility per group. Fair?
  - Disadvantaged group could have a lower maximum
  - Impacts of decisions may not be contained within groups
- People with the same estimates of $P[Y = 1|V = v]$ are treated the same. Fair?
  - Conditional probabilities change with variable selection
  - Estimates depend on sample, measurement, models
- Hmm, instead treat people the same if their true Y is the same?

Fairness flavor 1: equal prediction measures

- Treat people the same if their true Y is the same:
  - *Error rate balance* (Chouldechova, 2017): $D \perp A \mid Y$

Fairness flavor 2: equal decisions

- Forget Y. Why?
  - Y is very poorly measured
  - decisions are more *visible* than error rates (e.g. detention rates, lending rates)
- *Demographic parity*: $\mathrm{D} \perp A$

Fairness flavor 2: equal decisions

- *Unawareness/blindness*: $\delta(a, x_i) = \delta(a', x_i)$ for all i

Fairness flavor 3: metric fairness

- Related: people who are similar in x must be treated similarly
- More generally, a similarity metric can be *aware* of A:
- *Metric fairness* (Dwork et al., 2012): for every $v, v' \in \mathcal{V}$, their similarity implies similarity in decisions $|\delta(v) - \delta(v')| \leqslant m(v, v')$

Fairness flavor 3: metric fairness

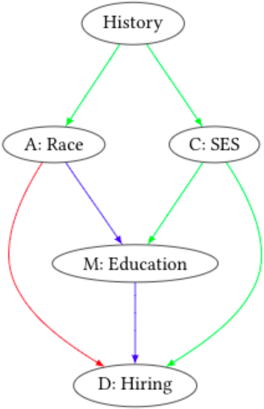
- How to define similarity $m(v, v')$...? Unclear.

Fairness flavor 4: causal

- Potential stuff again! a.k.a. counterfactuals
- $\mathrm{D}(\mathfrak{a})$ = decision if the person had their A set to $\mathfrak{a}$
- *Counterfactual Fairness*: $\mathrm{D}(\mathfrak{a}) = \mathrm{D}(\mathfrak{a}')$

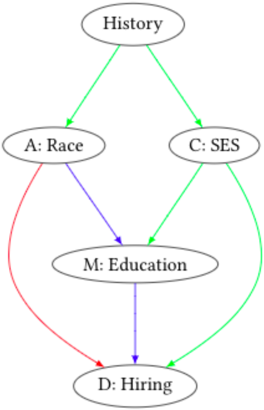
Fairness flavor 4: causal

- Instead of the *total* effect of A (e.g. race) on D (e.g. hiring), maybe some causal pathways from A are considered fair?
- Pearl (2009) defines causal graphs that encode conditional independence for counterfactuals:



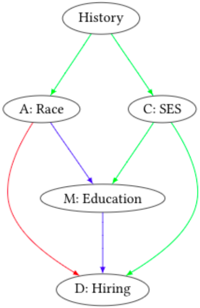
Fairness flavor 4: causal

- Zhang and Bareinboim (2018) decompose total disparity into disparities from each type of path: direct, indirect, and back-door



Fairness flavor 4: causal

- ML fairness definitions consider paths *from* A (e.g. race) (Nabi and Shpitser, 2018; Kilbertus et al., 2017)
- But what about back-door paths that contribute to disparity?
- Opinion: causal reasoning may be more useful to design interventions than to define fairness

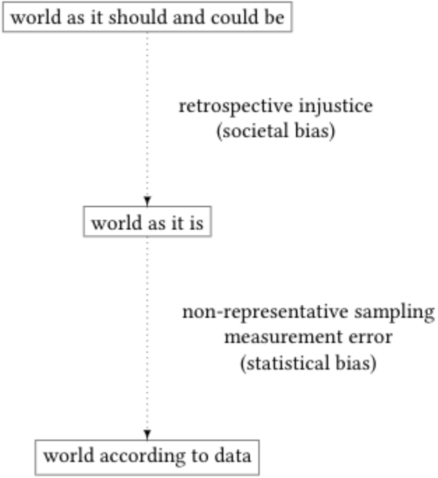


Confusing terminology

- Confusing:
  - $P[Y = 1|V = v]$ is called an individual's "true risk"
- But we have not measured all relevant attributes of an individual
- Instead:
  - individual i with measured variables v_i
  - $P[Y = 1|V = v]$ is a *conditional probability*

Confusing terminology

- "Biased data" collapses societal + statistical



"Conclusion"

- Neither maximizing a "utility function" (e.g. accuracy) nor satisfying a "fairness constraint" (e.g. demographic parity) guarantee social goals.
- But while data and mathematical formalization are far from saviors, they are not doomed to oppress. Purposeful alternatives are possible (Potash et al., 2015; Fussell, 2018).

Thank you!

arXiv.org > stat > arXiv:1811.07867

Searc
Help |

Statistics > Applications

Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, Kristian Lum

(Submitted on 19 Nov 2018 (v1), last revised 10 Jul 2019 (this version, v2))

A recent flurry of research activity has attempted to quantitatively define "fairness" for decisions based on statistical and machine learning (ML) predictions. The rapid growth of this new field has led to wildly inconsistent terminology and notation, presenting a serious challenge for cataloguing and comparing definitions. This paper attempts to bring much-needed order.
First, we explicate the various choices and assumptions made---often implicitly---to justify the use of prediction-based decisions. Next, we show how such choices and assumptions can raise concerns about fairness and we present a notationally consistent catalogue of fairness definitions from the ML literature. In doing so, we offer a concise reference for thinking through the choices, assumptions, and fairness considerations of prediction-based decision systems.

References I

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Fussell, S. (2018). The algorithm that could save vulnerable new yorkers from being forced out of their homes.

Harcourt, B. E. (2008). *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press.

Hu, L. and Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market.

Hu, L., Immorlica, N., and Vaughan, J. W. (2018). The disparate effects of strategic classification.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, pages 22–27.

References II

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. (2018). The social cost of strategic classification.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgenson, E., Mansour, R., and Ghani, R. (2015). Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2039–2047. ACM.

Zhang, J. and Bareinboim, E. (2018). Fairness in decision-making–the causal explanation formula. In *32nd AAAI Conference on Artificial Intelligence*.