# Designing equitable algorithms for finance and beyond

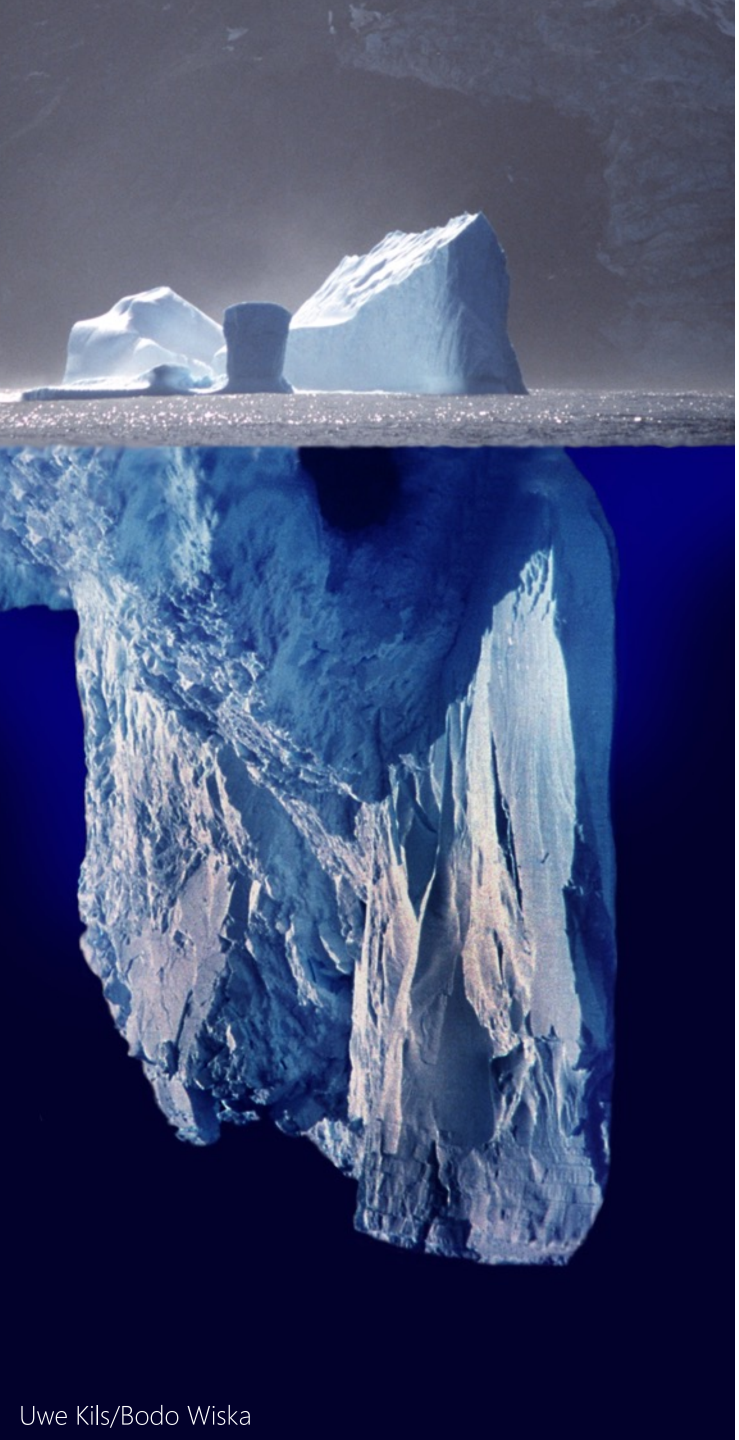
Sharad Goel, Harvard

Discussion

Jann Spiess, Stanford

ABFR Webinar

September 30, 2021



Uwe Kils/Bodo Wiska

# Just the tip of the iceberg

- The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning
- Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making

---

- Breaking Taboos in Fair Machine Learning: An Experimental Study. Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO 2021).
- A Large-scale Analysis of Racial Disparities in Police Stops Across the United States. Nature Human Behaviour, Vol. 4, 2020.
- Racial Disparities in Automated Speech Recognition. Proceedings of the National Academy of Sciences, Vol. 117, 2020.
- Algorithmic Decision Making and the Cost of Fairness. Proceedings of the 23rd Conference on Knowledge Discovery and Data Mining (KDD 2017).

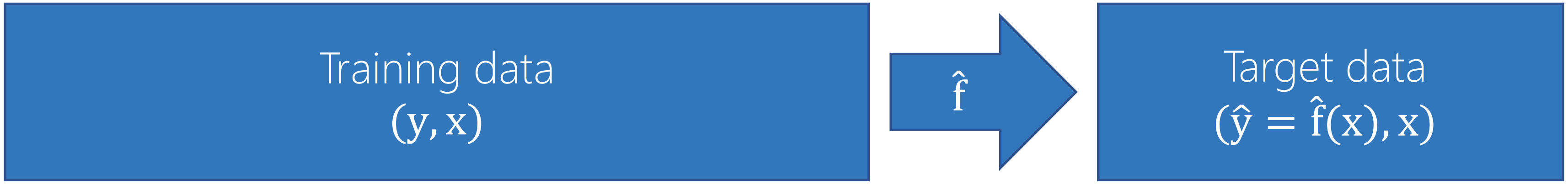
- Conceptual insight
- Algorithmic implementation
- Empirical evaluation
- Policy implications

# My comments

1. Place within broader inputs/outputs/consequences discourse
2. Summarize main ideas around deontological vs consequentialist
3. Mention some challenges
4. Move from design to oversight

# Broader input/output/consequence discourse

- High-stakes screening decisions performed by algorithms: Medical testing, hiring, lending
- Concerns around risk, fairness, bias (in general, welfare impact)
- Why not just inspect the prediction/classification function? After all, the structure is very similar to regression



Training data

$(y, x)$

$\hat{f}$

Target data

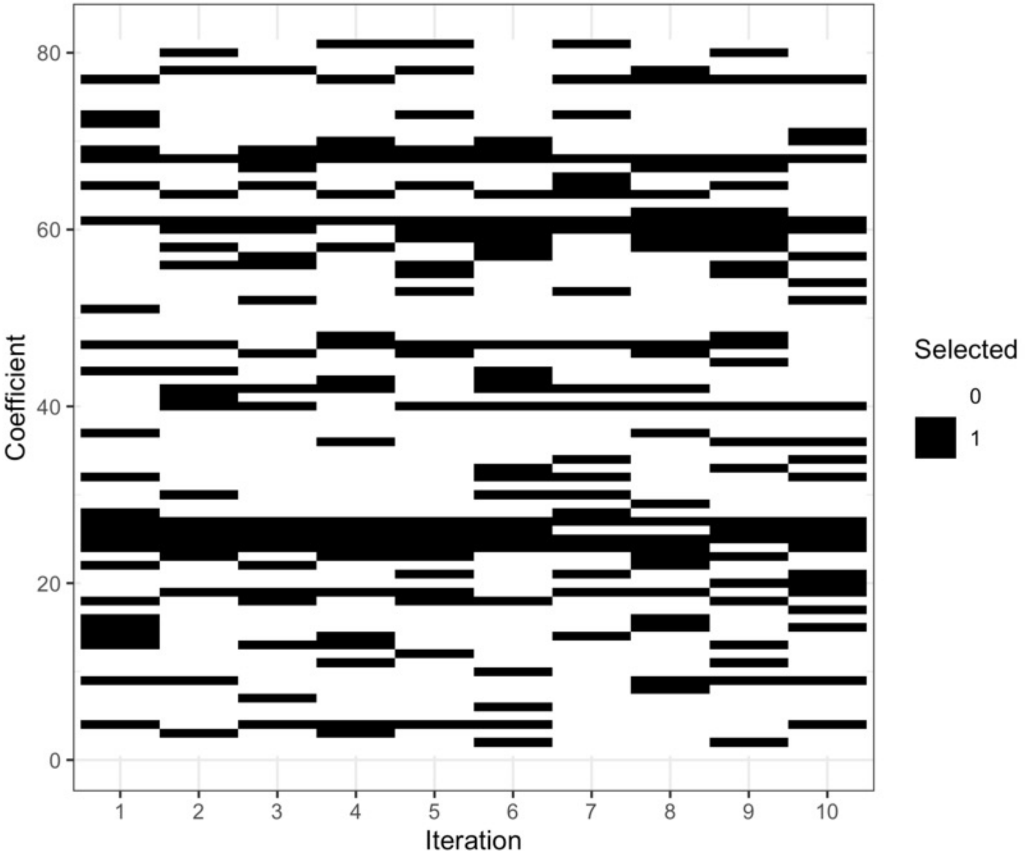
$(\hat{y} = \hat{f}(x), x)$

# From OLS…

$$
\begin{aligned}
P(default) = \alpha + \beta_1\ income + \beta_2\ age \\
+\beta_3\ education + \beta_4\ creditscore \\
+\beta_5\ x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80}
\end{aligned}
$$

# … to LASSO

$$P(default) = \alpha + \beta_1\ income + \beta_2\ age$$
$$+\beta_3\ education + \beta_4\ creditscore$$
$$+\beta_5\ x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80}$$



Gillis, T. B., & Spiess, J. L. (2019). Big data and discrimination. The University of Chicago Law Review, 86(2), 459-488.

$$P(default) = \alpha + \beta_1\ income + \beta_2\ age$$

$$+\beta_3\ education + \beta_4\ creditscore$$

$$+\beta_5\ x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80}$$

$$P(default) = \alpha + \beta_1\ income + \beta_2\ age$$
$$+\beta_3\ education + \beta_4\ creditscore$$
$$+\beta_5\ x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80}$$

$$P(default) = \alpha + \beta_1\, income \cancel{+\, \beta_2\, age}$$

$$+\beta_3\, education + \beta_4\, creditscore$$

$$+\beta_5\, x_5 + \cdots + \beta_{27} x_{27} + \cdots \beta_{80} x_{80}$$

$$age \approx f(income, creditscore, \ldots, x_{27}, \ldots)$$

# Broader input/output/consequence discourse

- High-stakes screening decisions performed by algorithms: Medical testing, hiring, lending
- Concerns around risk, fairness, bias (in general, welfare impact)
- Why not just inspect the prediction/classification function? After all, the structure is very similar to regression

- In high-dimensional data, representation not good way to think about statistical properties
- Simple input restrictions at best ineffective, can hurt
- Here: sidesteps input vs output and directly tackle properties of the decision, then go one step further and focus on consequences

# Evaluating decisions directly

- Anti-classification
  Gender does not affect decision
- Classification parity
  Similar error across groups
- Calibration
  Risk prediction means same thing across groups

- Precision
  Best possible prediction given available info

First main point: such fairness measures are not just incompatible (e.g. Chouldechova, 2017; Kleinberg, Mullainathan, and Raghavan, 2017), but also individually problematic

# Evaluating consequences (welfare maximization)

Second main point: directly optimize utility along Pareto frontier

$$\max \text{prediction fit} + \lambda \cdot \text{fairness}$$

- *I like this a lot* (not just as an economist) since it optimizes explicitly
  Rambachan, A., Kleinberg, J., Mullainathan, S., & Ludwig, J. (2020). An economic approach to regulating algorithms. NBER WP27111.
- *Do not see the bright distinction – many ways of measuring welfare*
- Does, of course, not tell us what the right measure of fairness is
  (and I wasn't quite clear how it addresses inframarginality in the example)
- While it does not make trade-offs go away, it makes them explicit
  (and allows their visual representation)
- *I also really appreciate that this is happening in application context*

# Challenges to analysis and implementation

$$\max \text{individual–level prediction fit} + \lambda \cdot \text{policy–level welfare outcome}$$

- Properties and procedure depend on our ability to diagnose welfare from the data

- Some unfair outcomes may relate to differential precision and data availability, so just predicting as well as possible does not alleviate issues
  Blattner, L., & Nelson, S. (2021). How Costly is Noise? Data and Disparities in Consumer Credit. arXiv preprint arXiv:2105.07554.
  *Here, fairness about investing in data, rather than how we use the one we have*
  *This also applies to the focus on prediction as a gold standard*

- The welfare-relevant consequences of interest may not be measured, measured with delay, or measured only with bias, may themselves be reflective of bad equilibria and institutional discrimination
  e.g. arrest, cost of incarceration, value of providing credit
  *Overall, getting to the right criteria here will be very context-specific,*
  *start with an acknowledgement that all we measure may reflect past discrimination*

- Optimization for training data, but we care about utility in deployment

# From design to oversight

$$\max \text{individual}-\text{level prediction fit} + \lambda \cdot \text{policy}-\text{level welfare outcome}$$

- **Preference may not be shared** between algorithm designer and regulator
- What if regulator cannot understand the complex classifier/has limited info?

- **Principal-agent perspective**
  Blattner, L., Nelson, S., & Spiess, J. (2021). Unpacking the Black Box: Regulating Algorithmic Decisions.

1. Regulator sets rules of the game, including possible restrictions to functions
2. Designer observes training data and chooses prediction/classifier
3. Outcomes/consequences get realized
4. Audit takes place

# From design to oversight

$$\max \text{individual}-\text{level prediction fit} + \lambda \cdot \text{policy}-\text{level welfare outcome}$$

- Ex-ante function restrictions inefficient
- Algorithmic audits can align incentives with the right design
  - Outcome tests based on realized consequences only
  - Best prediction explainers that minimize overall info loss
  - Targeted explainers that inspect misalignment
- Empirical case study in consumer lending using large-scale credit bureau data with race/ethnicity info to show relevance

Blattner, L., Nelson, S., & Spiess, J. (2021). Unpacking the Black Box: Regulating Algorithmic Decisions.

# My take-aways

- Explicitly model what we want and then optimize for it is not just a response to concerns about algorithms, but also big opportunity relative to the status quo (humans as ultimate biased black box)

- This work makes important and practically relevant contributions

- When we implement, need to solve questions that go beyond what optimization and statistics can do in training data:
  - Unmeasured or mismeasured consequences
  - Bad equilibria and engrained institutional discrimination
  - Fundamental uncertainty about deployment context
  - What is the goal and how is it enforced

- These require interdisciplinary work, *for which this is a great model*