Some ongoing debates in fair decision-making

Moritz Hardt

UC Berkeley

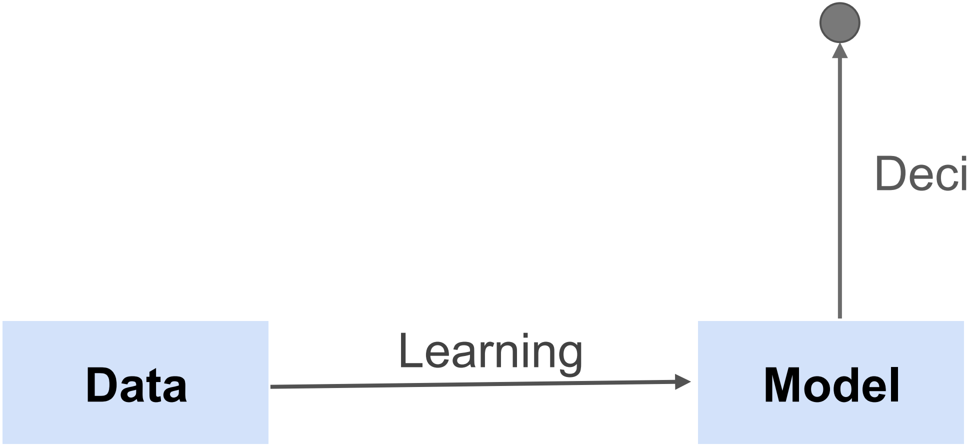
In this talk

Much current research engages with two common critiques of fair decision-making:

1. The decision-making framework is too narrow.
2. Standard criteria fail to take causality into account.

This list excludes other important ongoing research threads.

The standard view of learning and decision making



What’s missing?

CP

“[T]echnologies are developed and used within a particular social, economic, and political context. They arise out of a social structure, they are grafted on to it, and they may reinforce it or destroy it, often in ways that are neither foreseen nor foreseeable.”

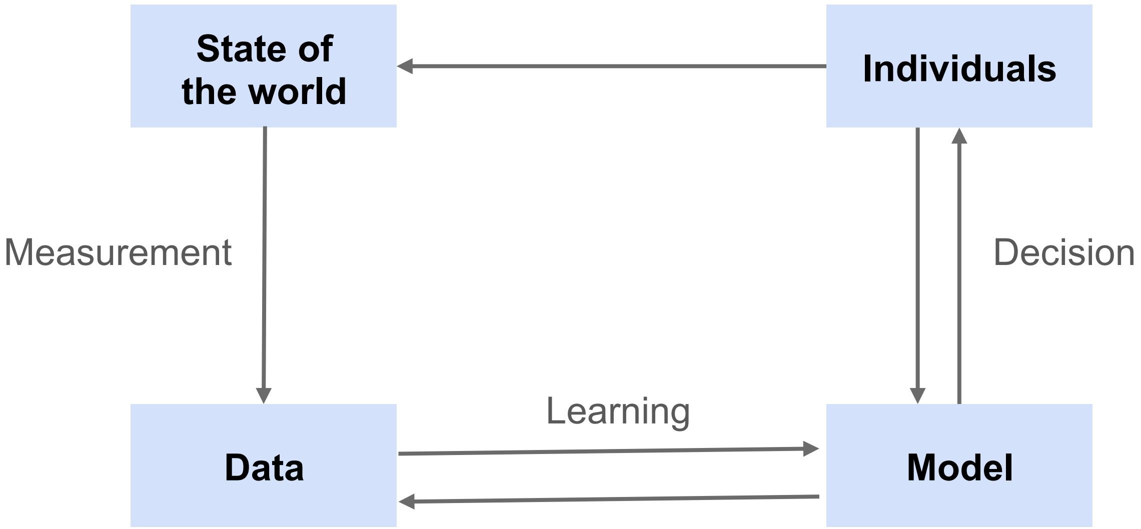
Ursula Franklin, 1989

CP

> *“[C]ontext is not a passive medium but a dynamic counterpart. The responses of people, individually, and collectively, and the responses of nature are often underrated in the formulation of plans and predictions.”*

Ursula Franklin, 1989

Social decisions in the real world



State of the world

Individuals

Measurement

Decision

Learning

Data

Model

Questions from a broader perspective

Long-term effects: What near-term interventions lead to long-term improvement?

Individual vs structural: What is the locus and scope to study discrimination?

Micro to Macro: Do fairness interventions at a micro level to fairness at a macro level?

Societal context: How do we formalize and take societal context into account?

Delayed impact of fair machine learning

Fairness research has produced numerous *fairness criteria*

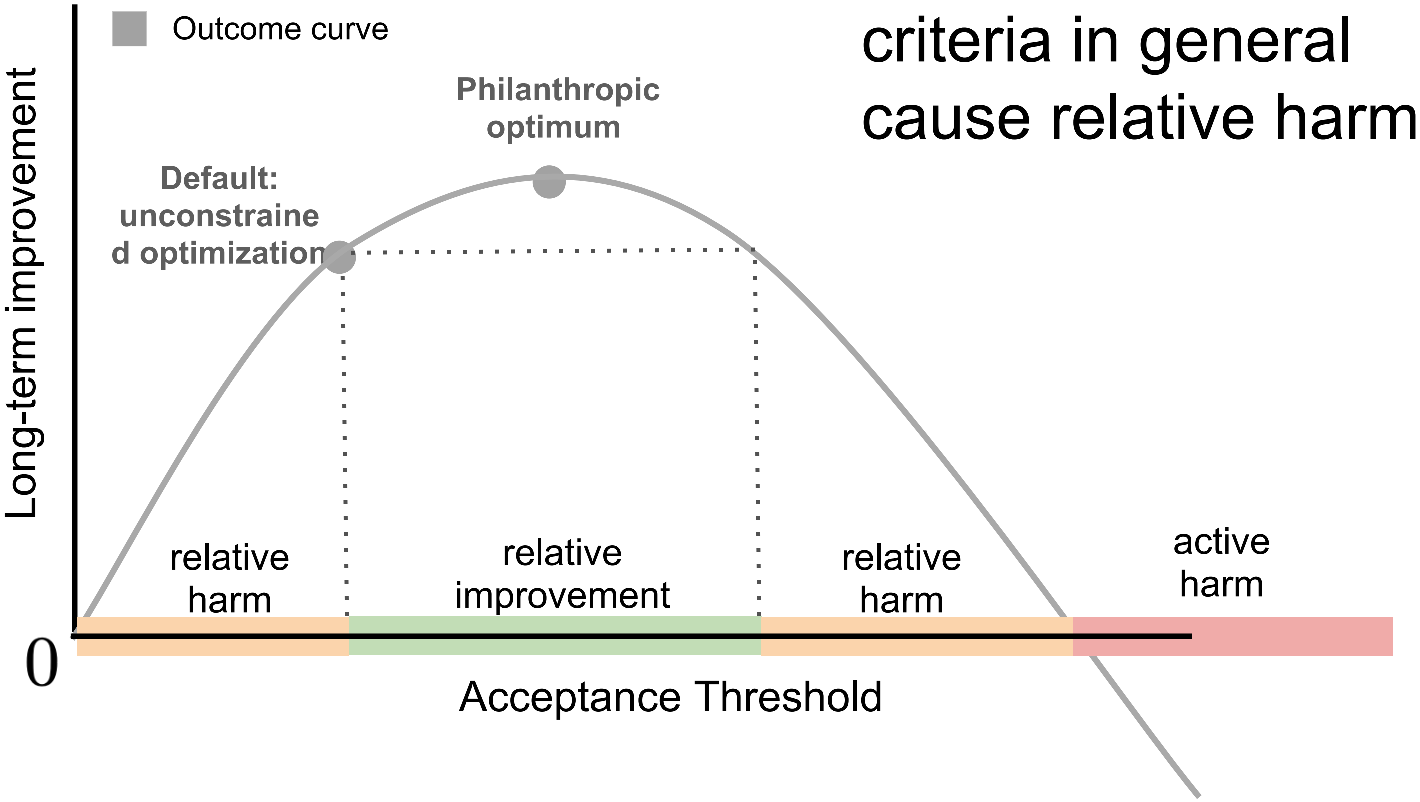
See: Fairness and Machine Learning. Barocas, *H*, Narayanan (fairmlbook.org)

Fairness criteria are typically conceptualized as near term interventions

Delayed impact: Do they promote long-term improvements for the groups they aim to protect?

Liu, Dean, Rolf, Simchowitz, *H* (2018)

Existing fairness criteria in general cause relative harm.



Fairness and dynamics

Much recent technical work on fairness in dynamic settings

Woulda, coulda, shoulda: Counterfactually-guided policy search, Buesing et al.; SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments, Bountouridis et al.; Counterfactual off-policy evaluation with Gumbel-max structural causal models, Oberst and Sontag; Causal modeling for fairness in dynamical systems, Creager et al.; Fairness without demographics in repeated loss minimization, Hashimoto et al.; A short-term intervention for long-term fairness in the labor market, Hu and Chen; Fairness in reinforcement learning Jabbari et al; Downstream effects of affirmative action Kannan et al.

Lots of room for pitfalls of systems modeling

Selbst et al. Fairness and Abstraction in Sociotechnical Systems; Baker. Model Metropolis.

Causality and fairness

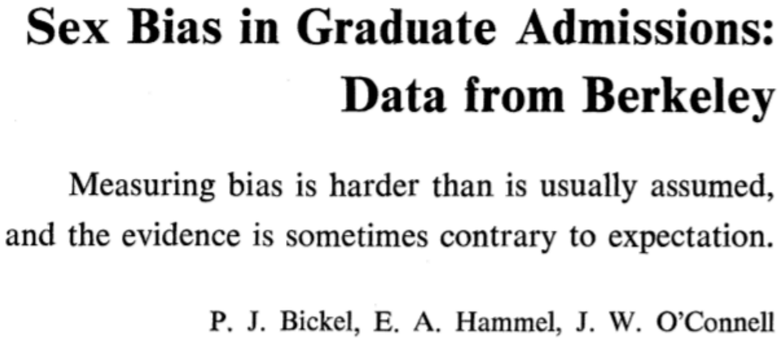
UC Berkeley grad admissions 1973

Data shows:

Male acceptance rate 44%

Female acceptance rate 30%

among top six departments

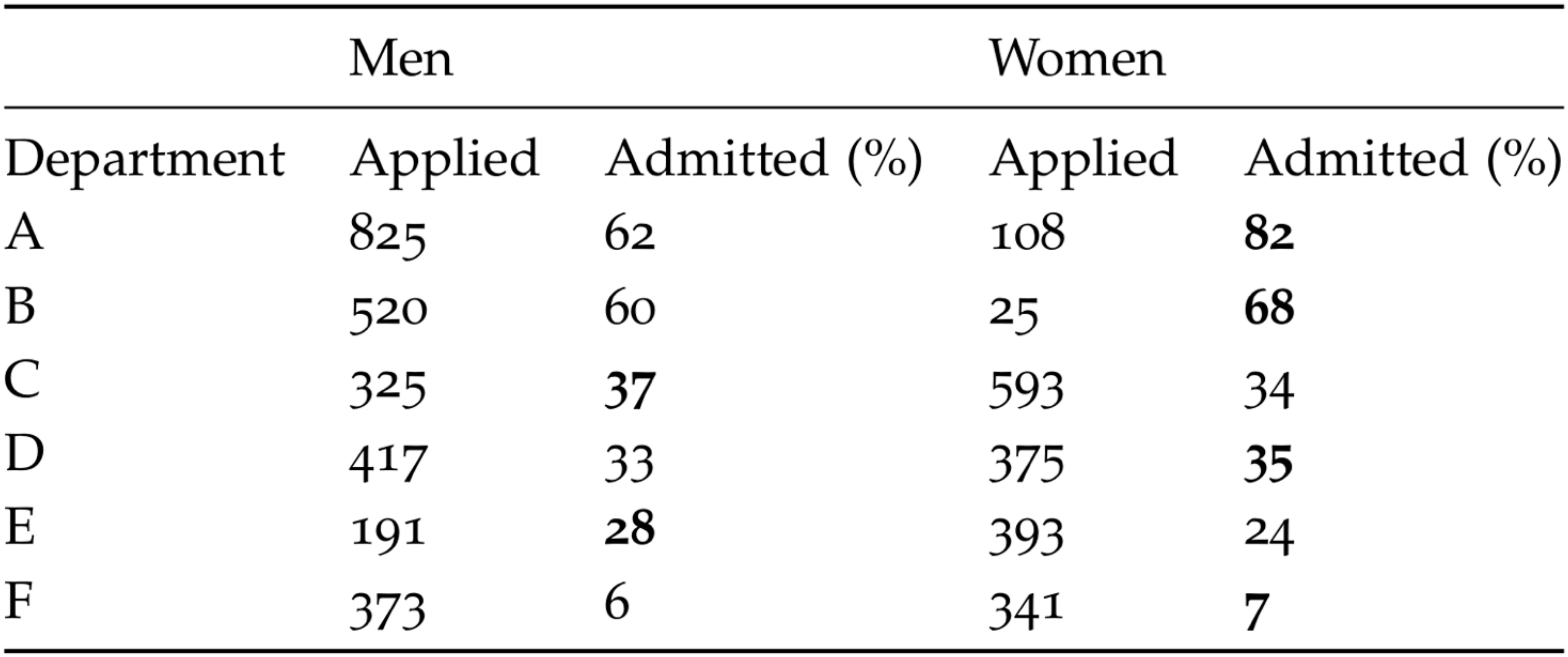


Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Analysis by department



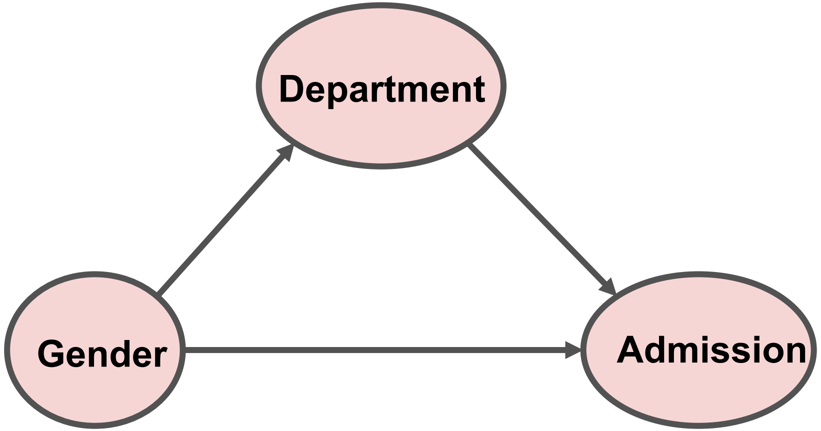
	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

Bickel’s explanation

The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

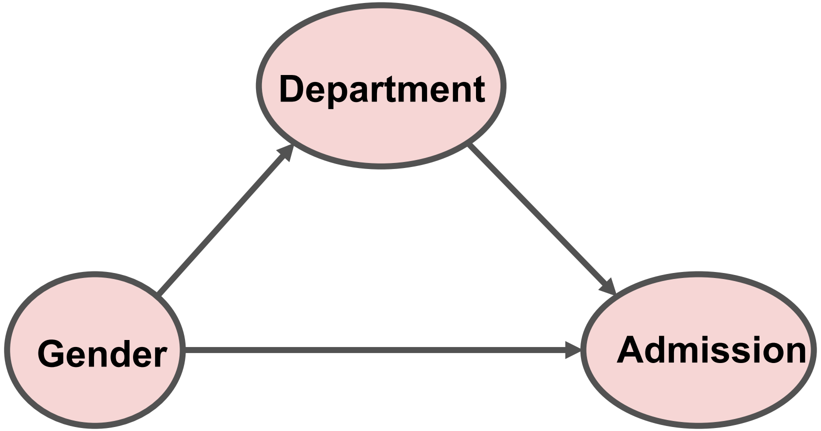
Bickel, Hammel, O’Connell (1975)

Pearl’s causal interpretation



Department choice **mediates** influence of Gender on Admission

Direct influence of Gender?



Pearl argues: Discrimination is the **direct effect** of Gender on Admission

Problems with discrimination as direct effect

Roughly captures *disparate treatment* doctrine in the law and inherits its pitfalls

In particular, defeated by proxy discrimination

Indirect paths can encode discrimination

E.g., women choose not to apply to department X because it compensates them in a blatantly unfair way

What then is the right causal fairness criterion?

Ongoing debate

Causality clarifies issues of confounding, mediation etc., but it does not on its own resolve the normative debate of what we ought to consider fair

Barocas, Hardt, Narayanan. Fairness and Machine Learning. Chapter 4 (Causality); Avoiding discrimination through causal reasoning Kilbertus et al.; Counterfactual fairness, Kusner et al.; To predict and serve? Lum and Isaac; Residual unfairness in fair machine learning from prejudiced data, Zhou and Kallus;

Testing discrimination in practice (audits and experiments)

See Chapter 5.

Robust critique of including *race* and *gender* in causal models and counterfactuals

Kohler-Hausmann. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination;
Hu. Disparate causes. Part 1 and Part 2; Hu and Kohler-Hausmann 2020
Not to be confused with debate around manipulability in causality.

Thank you