Fairness and Machine Learning: Limitations and Opportunities

Moritz Hardt
UC Berkeley

Where to start?

We live in a world of pervasive inequality, oppression, and discrimination.

As we use machine learning to formalize, scale, accelerate processes in this world, we run danger of perpetuating existing patterns of injustice.

But there's also a (somewhat fragile) opportunity to revisit decision making in various domains and reform existing processes for the better.

Important work to start with

Ruha Benjamin. Race After Technology: Abolitionist Tools for the New Jim Code

Meredith Broussard. Artificial Unintelligence: How Computers Misunderstand the World

Virginia Eubanks. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor

Safiya Noble. Algorithms of Oppression: How Search Engines Reinforce Racism

Cathy O'Neil. Weapons of math destruction

Joy Boulamwini, Kate Crawford, Timnit Gebru, Latanya Sweeney, Meredith Whitaker and many others.

“the New Jim Code”: *the employment of new technologies that reflect and reproduce existing inequities but that are promoted and perceived as more objective or progressive than the discriminatory systems of a previous era.*

- Ruha Benjamin

Focus for this tutorial

Discrimination in consequential decision making settings

- This excludes many other forms of injustice (and even unfairness)

US centric perspective (insofar as the examples and legal backdrop go)

Formal models and frameworks

- This is not meant to decenter the scholarship just mentioned
- This decidedly leaves room for non-technical interventions

Is discrimination not the point of machine learning?

Our concern is with *unjustified basis for differentiation*

- *Practical irrelevance*
  - *Sexual orientation in employment decisions*
- *Moral irrelevance*
  - *Disability status in hiring decisions*

Discrimination is *not* a general concept

Domain specific:

Concerned with important opportunities that affect people's lives

Group specific:

Concerned with socially salient categories that have served as the basis for unjustified and systematically adverse treatment in the past

Regulated domains (based on US law)

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **‘Public Accommodation’** (Civil Rights Act of 1964)

Extends to *marketing and advertising*; not limited to final decision

This list sets aside complex web of laws that regulates the government

Legally recognized ‘protected classes’ in the US

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Courts & Law

Supreme Court says gay, transgender workers protected by federal law forbidding discrimination

Source: Washington Post
June 15, 2020

Two legal doctrines in the US

Disparate treatment

Purposeful consideration of group membership

Intentional discrimination without consideration of group membership

Goal: Procedural fairness

Disparate impact

Avoidable or unjustified harm, possibly indirect

Goal: Distributive justice, minimize differences in outcomes

Some well-recognized tension between the two.

Some caveats about the law

Anti-discrimination law does not reflect one moral theory

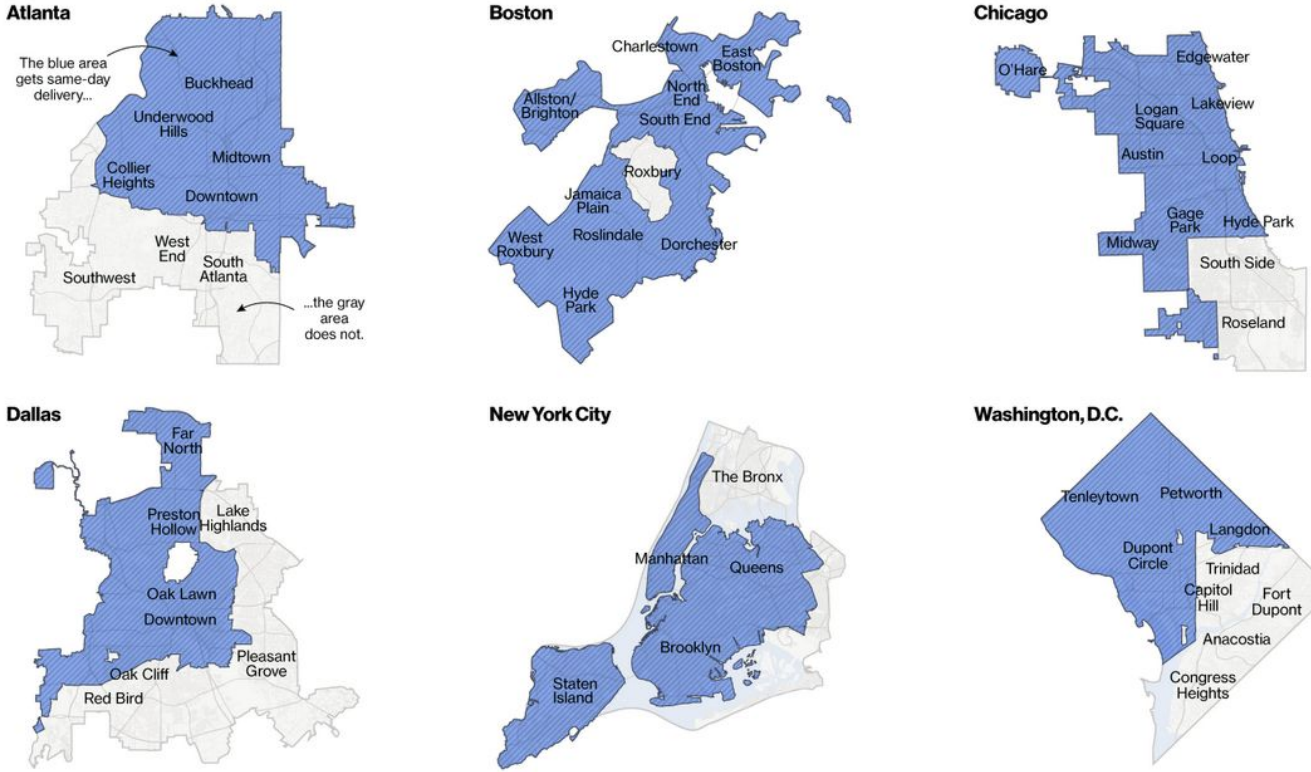
Legislations often were responses to civil rights movements, each hard fought through decades of activism

The law does not give us a “fairness definition” that we could readily formalize and operationalize

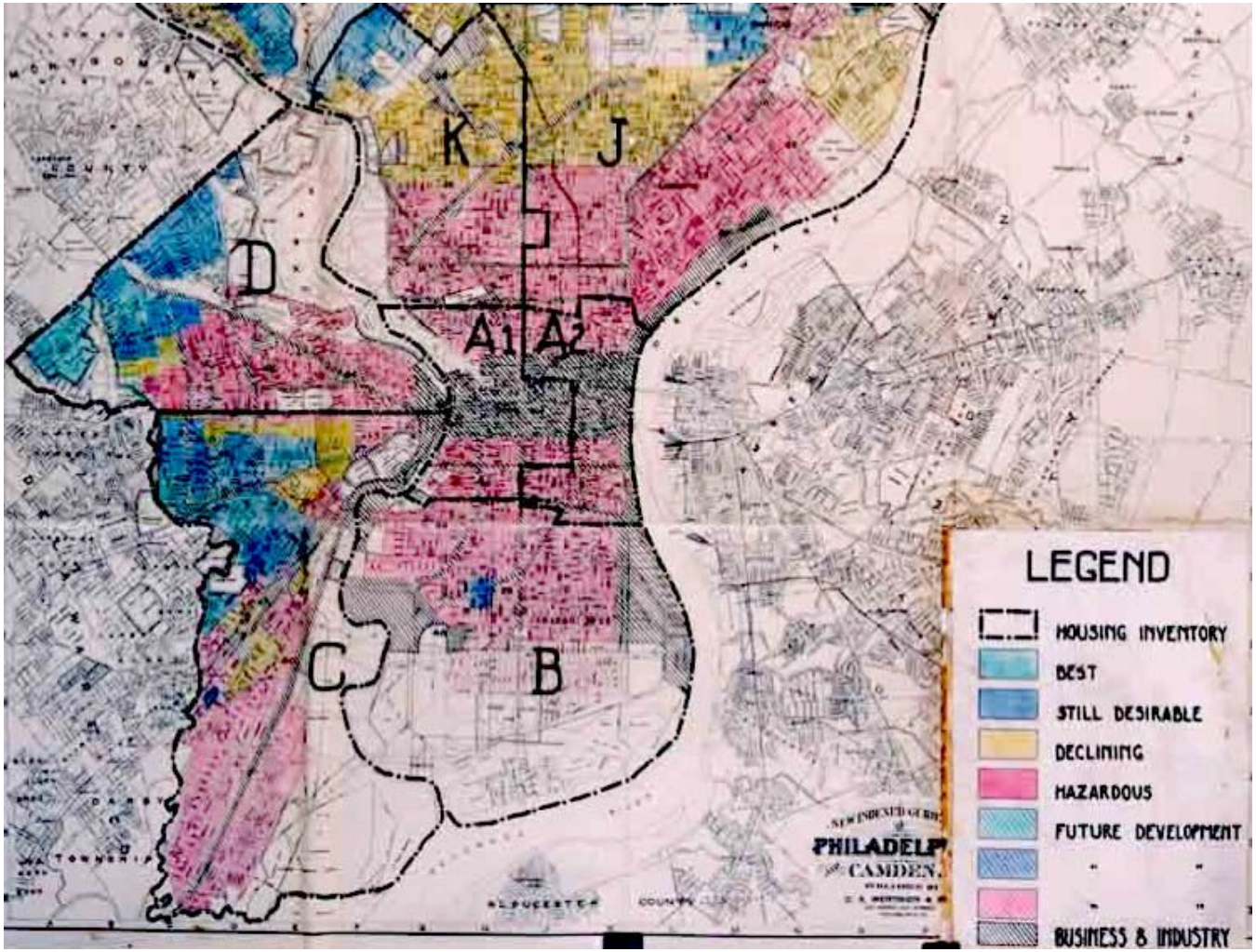
The failure of *fairness through unawareness*

Removing (or not including) “sensitive attributes” is no cure for fairness concerns and can exacerbate them.

Amazon same-day delivery coverage



Source: https://www.bloomberg.com/graphics/2016-amazon-same-day/


K
J
D
A1
A2
C
B
LEGEND
HOUSING INVENTORY
BEST
STILL DESIRABLE
DECLINING
HAZARDOUS
FUTURE DEVELOPMENT
BUSINESS & INDUSTRY

The failure of fairness through *unawareness*

Perhaps Amazon was just predicting *number of purchases*, which correlates with affluence, which correlates with race in the United States. *Amazon almost certainly did not look at their customers' race when they built this product.*

“We don’t consider that in our data” is *never* a valid argument.

So what should we do instead?

Overview

Part I (today): From a narrower perspective

Fairness criteria in classification

Part II (Thu): Toward a broader perspective

Causal models of decision-making settings

Dynamic models of socio-technical systems

Part I

Formal work on fairness in classification and decision-making

Pioneering work in educational testing (Cleary 1968) and economics (Becker 1957, Phelps 1972, Arrow 1973) on the heels of civil rights movement.

Computer science: Mostly post 2010, explosive increase in work since 2016

Why today? Urgency, scale, reach, and impact of algorithmic decisions

Machine learning fuels adoption and motivates some new technical problems, but also forces us to revisit fundamental normative questions

Formal prediction and decision making setting

Data described by covariates $\boldsymbol{X}$

Outcome variable $\boldsymbol{Y}$ (often binary, sometimes called *target variable*)

Our goal is to *predict* $\boldsymbol{Y}$ from $\boldsymbol{X}$

Use supervised machine learning to produce a score function $\boldsymbol{R = r(X)}$

Make binary decisions according to threshold rule $\boldsymbol{D} = \mathbf{1}\{\boldsymbol{R} > t\}$

Note: Think of these as random variables in the same probability space.

Where do score functions come from

Score ***R*** could be:

- Based on parametric model of the data **(*X, Y*)**, e.g. *likelihood ratio test*
- Non-parametric score, such as, Bayes optimal score **$R = E[\, Y \mid X \,]$**
- Most commonly, learned from labeled data using supervised learning

Decision theory 101

Outcome Y	Decision *D* = 0	Decision *D* = 1
0	*True negative*	*False positive*
1	*False negative*	*True positive*

True positive rate = Pr[***D*** = 1 | ***Y*** = 1]

False positive rate = Pr[***D*** = 1 | ***Y*** = 0]

True negative rate = Pr[***D*** = 0 | ***Y*** = 0]

False negative rate = Pr[***D*** = 0 | ***Y*** = 1]

Statistical fairness criteria

Introduce additional random variable ***A*** encoding membership status in a protected class

Equalize different statistical quantities involving group membership ***A***

Idea dates back at least to the 1960s with work of Anne Cleary about group differences in educational testing*

We'll review *three* common criteria

* See Hutchinson, Mitchell (2018).

Equalizing acceptance rate

Equal positive rate: For any two groups *a, b*, require
Pr[***D*** = 1 | ***A*** = *a*] = Pr[***D*** = 1 | ***A*** = *b*]

“Acceptance rate” equal in all groups

Generalization: Require ***D*** to be independent of ***A*** (Independence)

All sorts of variants, relaxations, equivalent formulations

Why this does not rule out *unfair* practices

One unfair situation: Make good/informed decisions in one group, poor/arbitrary decisions in other groups. Equalize positive rate.

This can happen on its own if we have less data or poor data in one group.

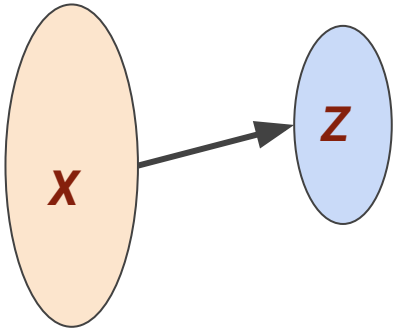
Example: Old *Framingham risk score* for coronary heart disease was created on cohort of white men, then used for other patients.

A *positive* call could be a false positive or a true positive. Moral intuition: You shouldn't get to match true positives in one group with false positives in another.

Achieving independence through representation learning

Lots of work out there on “fair representation” starting with work by Zemel et al. (2015).

General idea: Use deep learning tricks, such as adversarial learning, to train a representation of the data that is independent of group membership ***A***, while representing original data as well as possible.



$Z \perp A$

Equalizing error rates

For any two groups *a*, *b*, require

$\Pr[\boldsymbol{D} = 1 \mid \boldsymbol{Y} = 0, \boldsymbol{A} = a] = \Pr[\boldsymbol{D} = 1 \mid \boldsymbol{Y} = 0, \boldsymbol{A} = b]$ (equal false positive rate)

$\Pr[\boldsymbol{D} = 0 \mid \boldsymbol{Y} = 1, \boldsymbol{A} = a] = \Pr[\boldsymbol{D} = 0 \mid \boldsymbol{Y} = 1, \boldsymbol{A} = b]$ (equal false negative rate)

Generalization: Require ***D*** to be independent of ***A*** *given* ***Y***

Also makes sense for score: Require ***R*** to be independent of ***A*** *given* ***Y***

Error rate parity is a *post-hoc* criterion

At decision time, the decision maker doesn't know who is a positive/negative instance

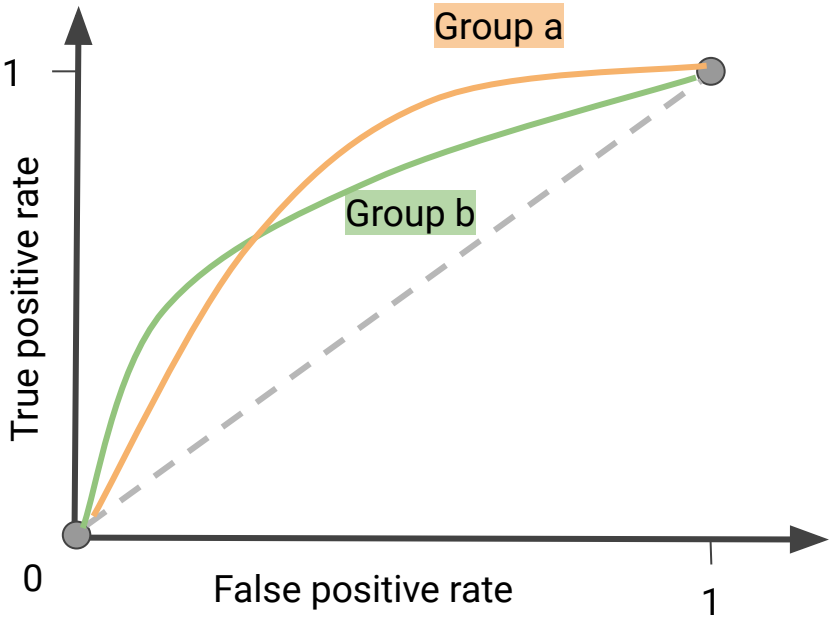
In hindsight, somebody can collect a group of positive instances and a group of negative instances and check how they were classified.

Group differences in this kind of post-hoc "audit" often strike people as unfair.

Interpretation in terms of ROC curve

Suppose ***D*** is threshold of a score ***R***

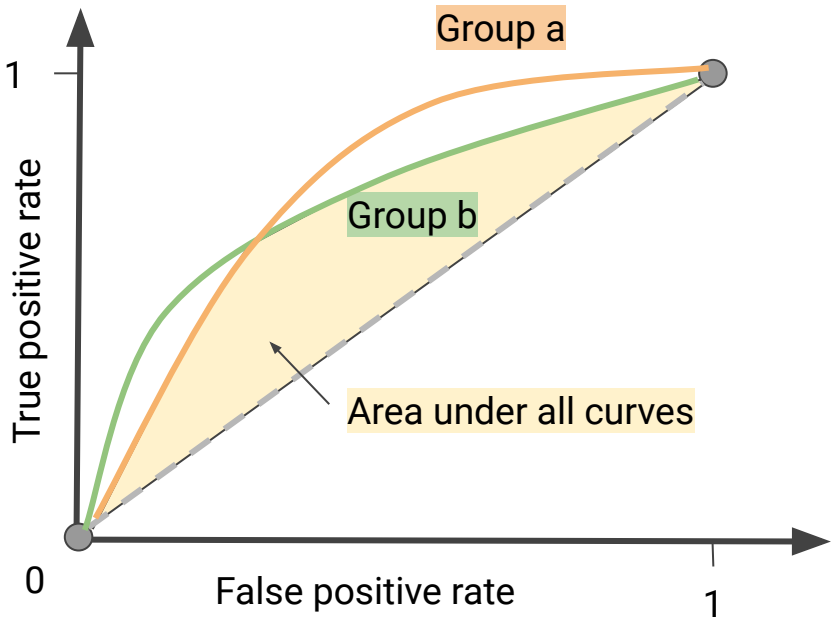
Error rate parity implies that ROC curve of score conditional on group must be under all curves.



Interpretation in terms of ROC curve

Suppose ***D*** is threshold of a score ***R***

Error rate parity implies that ROC curve of score conditional on group must be under all curves.

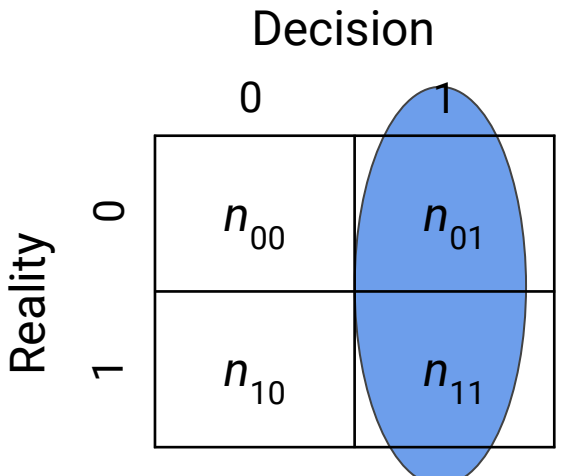


Column-wise criteria?

We could equalize expressions of the form
Pr[***Y*** = y | ***D*** = d, ***A*** = *a*]

These are called “column-wise” rates, i.e., false omission and false discovery rate.

Nothing wrong with this, but something closely related but different is more common.



Calibration

A score ***R*** is calibrated if: $\Pr[\boldsymbol{Y} = 1 \mid \boldsymbol{R} = r] = r$

“You can pretend score is a probability” - although it may not actually be one!

Score value *r* corresponds to positive outcome rate *r*

Calibration by group: $\Pr[\boldsymbol{Y} = y \mid \boldsymbol{R} = r, \boldsymbol{A} = a] = r$

Follows from: ***Y*** independent of ***A*** conditional on ***R***

Calibration is an *a priori* guarantee

The decision maker sees the score value *r* and knows based on this what the frequency of positive outcomes is.

E.g., score 0.8 means 80% rate of heart failure on average over people who receive score 0.8.

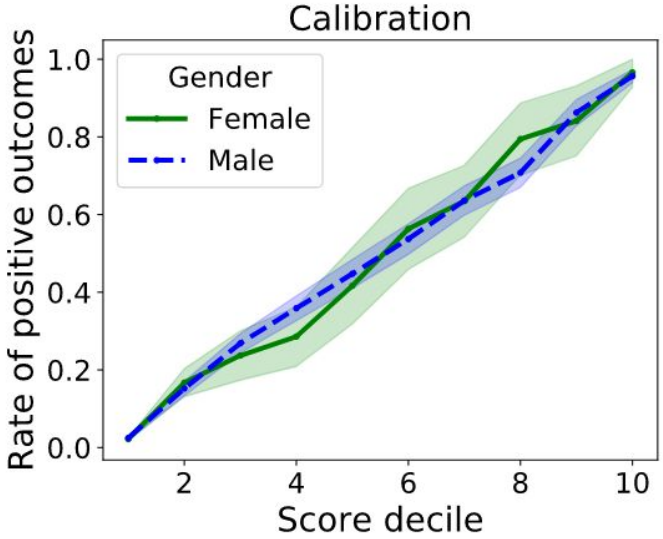
This guarantee (usually) does not hold at the individual level, e.g., “Mary’s individual risk of heart failure is 80%.”

Group calibration often follows from unconstrained learning

Informal theorem: Under reasonable conditions, the deviation from satisfying group calibration is upper bounded by the excess risk of the learned score relative to the Bayes optimal score function.

See Liu, Simchowitz, H (2019)

In other words, you shouldn't be surprised to see calibration follow approximately from unconstrained supervised learning.



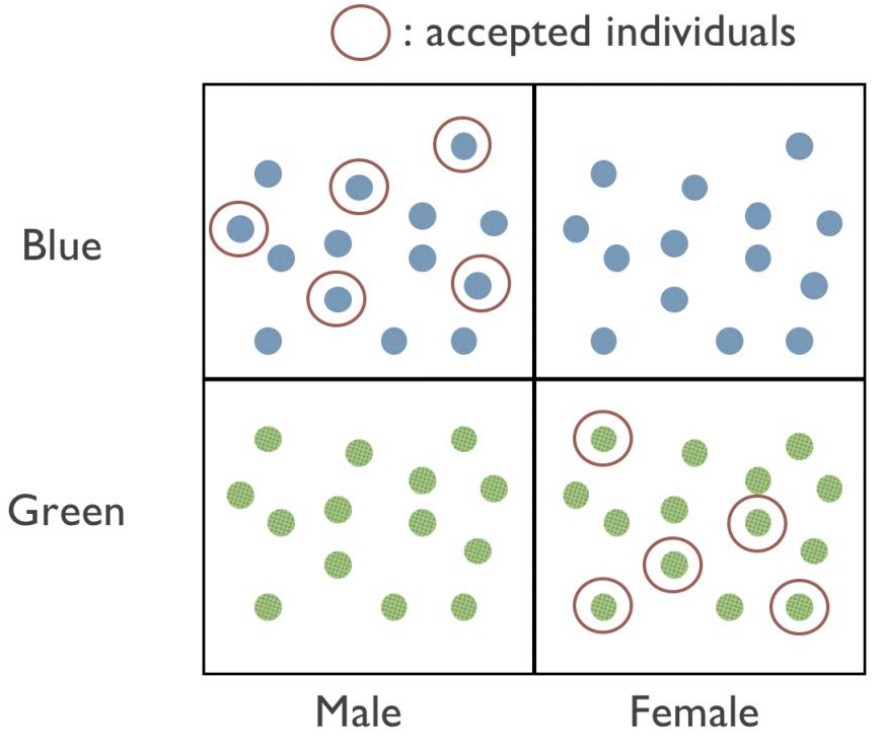
Example calibration of unconstrained learning on UCI adult data set.

Subgroup fairness

Ensuring fairness criteria between two groups can lead to violations within groups.

This motivated work on ensuring **subgroup fairness**.

See Kearns, Neel, Roth, Wu (2018), and Hébert-Johnson, Kim, Reingold, Rothblum (2018).



Illustrative example
from Kearns et al. (2018)

Recap

We saw three criteria:

- ***R*** independent of ***A*** (implies equal acceptance rate)
- ***R*** independent of ***A*** conditional on ***Y*** (implies equal error rates)
- ***Y*** independent of ***A*** conditional on ***R*** (implies calibration by group)

Can we have them all?

Incompatibility results

Informal theorem: Any two of these criteria are mutually exclusive in general.

Error rate parity vs calibration

Theorem:

1. Assume unequal base rates: $\Pr[\boldsymbol{Y} = 1 \mid \boldsymbol{A} = a] \neq \Pr[\boldsymbol{Y} = 1 \mid \boldsymbol{A} = b]$
2. Assume imperfect decision rule: $\boldsymbol{D}$ has nonzero error rates

Then, calibration by group implies that error rate parity fails.

Related result due to Chouldechova (2016), Kleinberg, Mullainathan, Raghavan (2017)

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Essence of COMPAS debate

There's a risk score used, called COMPAS, used by many jurisdictions in the United States to assess "risk of recidivism". Judges may detain defendant in part based on this score.

ProPublica: Black defendants face higher false positive rate, i.e., more Black defendants labeled "high risk" end up ***not*** committing a crime upon release than among Whites labeled "high risk"

COMPAS maker Northpointe: But our scores are calibrated by group and Black defendants have a higher recidivism rate! Hence, this is unavoidable.

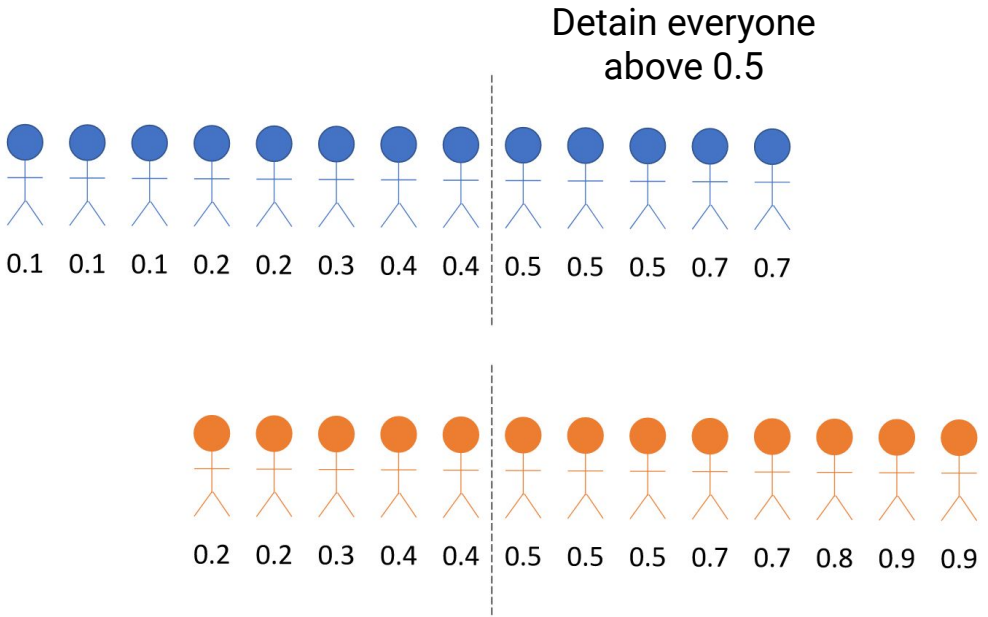
A first word of caution about COMPAS debate

Neither error rate parity nor calibration rule out blatantly unfair practices.

What's fair in criminal justice is not settled by appeal to one or the other criterion.

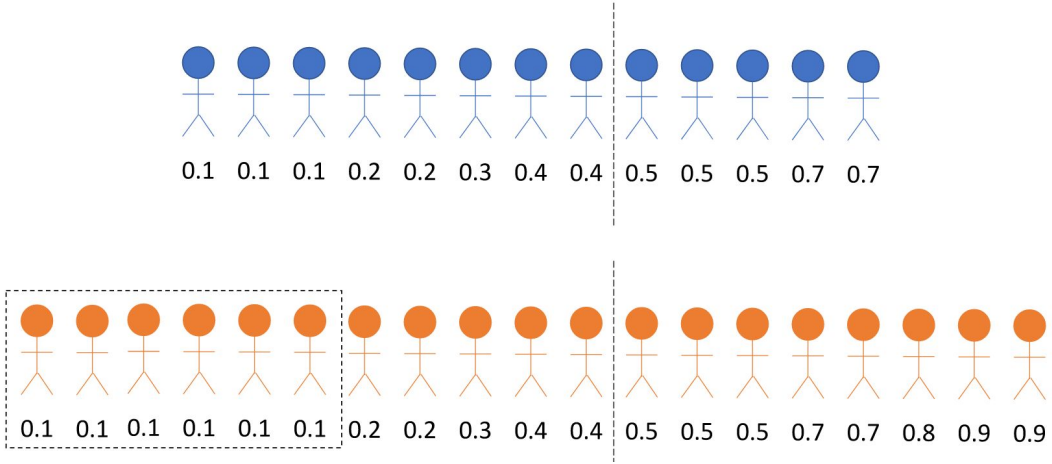
These properties are not meant to be "fairness certificates".

Consider these two groups



Detention rate	False pos. rate
38%	25%
61%	42%

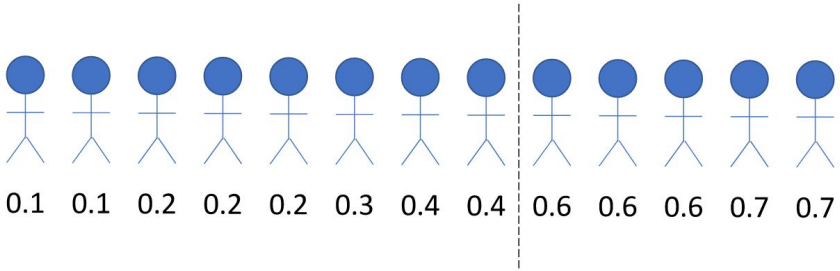
Equalizing rates may lead to undesired outcomes



Arrest more low risk individuals in orange group!

Detention rate	False pos. rate
38%	25%
~~61%~~ 42%	~~42%~~ 26%

An issue with calibration

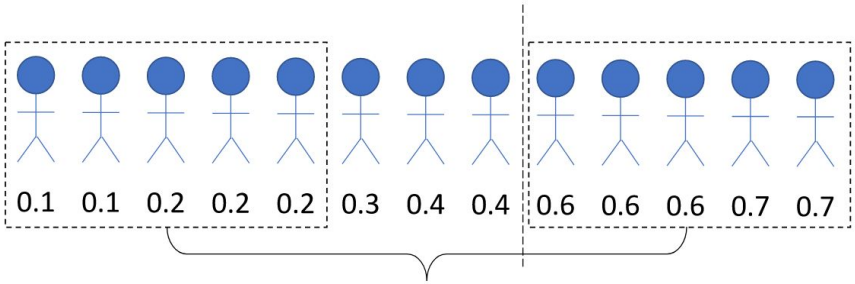


True probabilities of reoffending
(hypothetically, suppose we know them!)

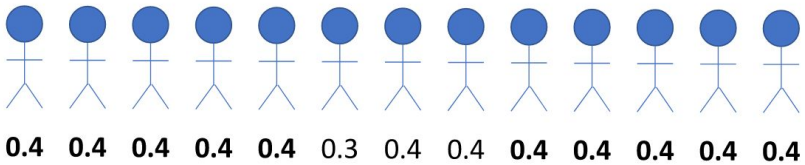
Detain everyone
above 0.5

Examples from: [Corbett-Davies, Pierson, Feller, Goel, Huq (2017)]

An issue with calibration



Average probability of re-offense is 0.4 in this subgroup



Calibrated new scores

No one is detained!

Is *prediction* too narrow a perspective?

Scholarly debate around COMPAS was largely about **tension between fairness criteria**.

Some rightfully point out **data and measurement problems**
(e.g., policing patterns influence variables such as criminal history and recidivism)

When is the issue not *how* we predict but *that* we predict?

Failure to appear in court

One approach: Predict failure to appear, jail if risk is high.

Alternative: Recognize that people fail to appear in court due to lack of child care and transportation, work schedules, or too many court appointments. Implement steps to mitigate these issues.

Alternative is part of the Harris County Lawsuit settlement: *"require Harris County to provide free child care at courthouses, develop a two-way communication system between courts and defendants, give cell phones to poor defendants and pay for public transit or ride share services for defendants without access to transportation to court." (Source: Houston Chronicle, April 2019)*

Toward a broader perspective

Statistical fairness criteria take data generating distribution as given and work with nothing but the joint statistics of ***(X, Y, R, A).***

If this statistical perspective is too narrow, how do we take salient social facts and context into account?

This will be the subject of Part II on Thursday.

Wrapping up

Fairness through unawareness *fails*!

Violations of fairness criteria trigger valid moral intuitions, and can surface normative questions about decision-making, as well as trade-offs and tensions between different interpretations of fairness.

However, statistical fairness criteria on their own cannot be a “proof of fairness”.

Nor are fairness criteria on their own a good objective function.

Background reading

A textbook in progress (mostly available online):

Barocas, H, Narayanan. Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org

Today's lecture roughly corresponds to Chapters 1 & 2.

Part II

Last time: Statistical fairness criteria

Covariates ***X***, outcome ***Y***, risk score ***R***, *status* ***A*** in a designated protected group

Fairness criteria give different answers to *"equality of what?"* question

Three common criteria:

- **Demographic parity:** Equal acceptance rates in all groups
  - Follows from: ***R*** independent of ***A***
- **Error rate parity:** Equal true/false positive rates in all groups
  - Follows from: ***R*** independent of ***A*** given ***Y***
- **Predictive parity:** Equal positive outcome rate given score in all groups
  - Follows from: ***Y*** independent of ***A*** given ***R***

Is *prediction* too narrow a perspective?

Scholarly debate around COMPAS was largely about **tension between fairness criteria**.

Some rightfully point out **data and measurement problems**
(e.g., policing patterns influence variables such as criminal history and recidivism)

When is the issue not *how* we predict but *that* we predict?

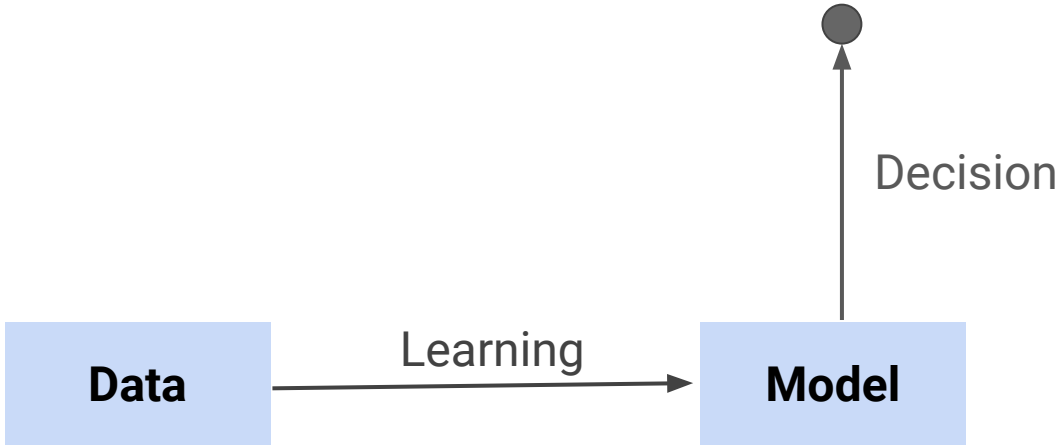
Failure to appear in court

One approach: Predict failure to appear, jail if risk is high.

Alternative: Recognize that people fail to appear in court due to lack of child care and transportation, work schedules, or too many court appointments. Implement steps to mitigate these issues.

Alternative is part of the Harris County Lawsuit settlement: *"require Harris County to provide free child care at courthouses, develop a two-way communication system between courts and defendants, give cell phones to poor defendants and pay for public transit or ride share services for defendants without access to transportation to court." (Source: Houston Chronicle, April 2019)*

The standard view of learning and decision making



What's missing?

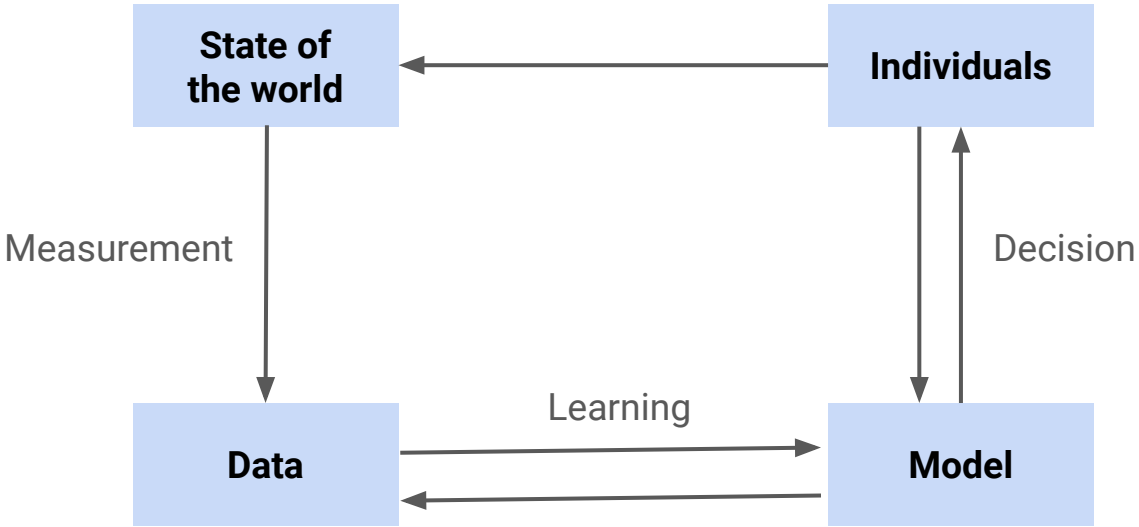
“[T]echnologies are developed and used within a particular social, economic, and political context. They arise out of a social structure, they are grafted on to it, and they may reinforce it or destroy it, often in ways that are neither foreseen nor foreseeable.”

Ursula Franklin, 1989

“[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



Two directions motivated by a broader perspective

Can we capture social context through causal models? What can causality say about questions of discrimination and fairness?

Can we formalize sociotechnical systems using dynamic models? How can use this to reason about feedback and long-term effects?

There are other notable approaches to take social context into account, e.g., the individual similarity measure in Dwork, H, Pitassi, Reingold, Zemel (2012)

Causal modeling and fairness

Loosely based on Chapter 4 in fairmlbook.org

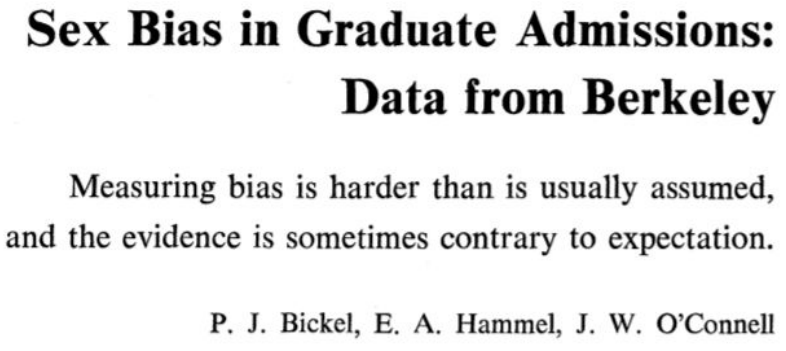
UC Berkeley grad admissions 1973

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

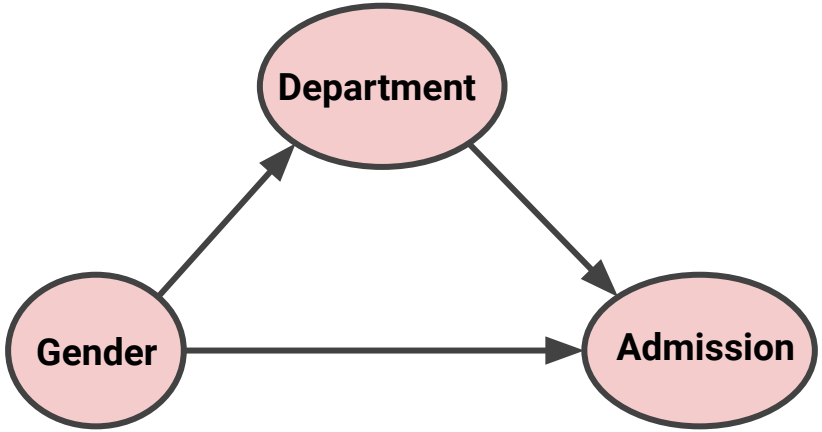
	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

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.

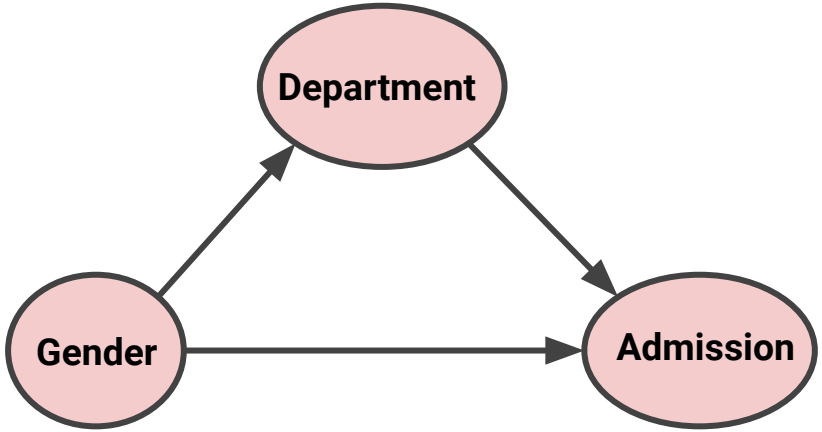
Bickel, Hammel, O'Connell (1975)

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

Normative problems with discrimination as direct effect

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

In particular, defeated by indirect discrimination

Indirect paths can encode discrimination: E.g., women choose not to apply to a department because the compensation is lower for women, or because the department policies are harmful to women

Why do we think that department choice as a mediator disarms concerns of discrimination?

Other causal fairness criteria?

If not direct effects, what else can we do?

There's been much work on causal fairness criteria and path effects. See survey by Chiappa and Isaac (2019), as well as Chapter 4 at fairmlbook.org for discussion.

But there's another problem I want to focus on in this talk.

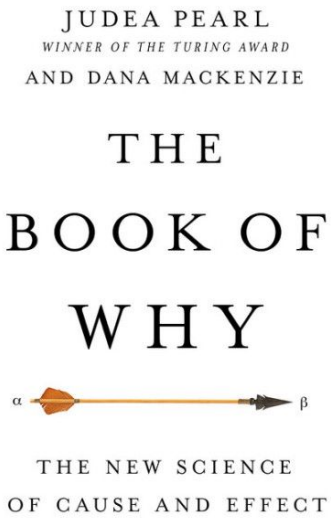
From “Book of Why” on Berkeley admissions

“We get the direct effect of X on Y when we **‘wiggle’** X [here, gender] without allowing M [here, department] to change.” p. 317

“Hold constant the variable Department and then **tweak** the variable Gender” p. 315

“**force** everybody to apply to the history department [...] randomly assign some people to **report their sex** as male” p. 317

“would be admitted if she **faked her sex** to read ‘male’” p. 318



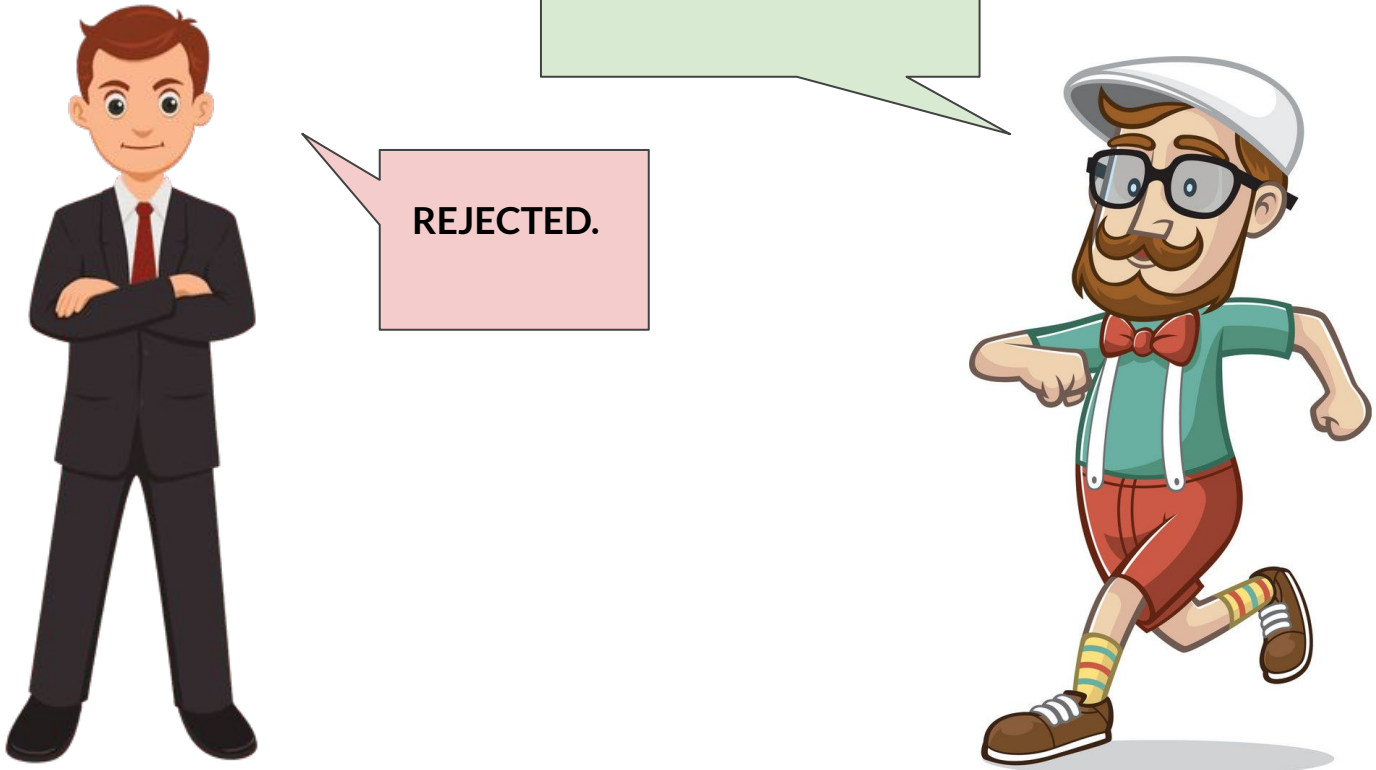
Two neglected problems in causal modeling

The ontological problem:
What is the thing a node in a causal graph references?

The epistemological problem:
How do we know facts about this thing?

I will focus on the former, illustrate it with a lighthearted example, and then argue why this is serious a problem we cannot ignore (be it theoretically or practically).

How do you do, good Sir? I'm here to apply for the job.



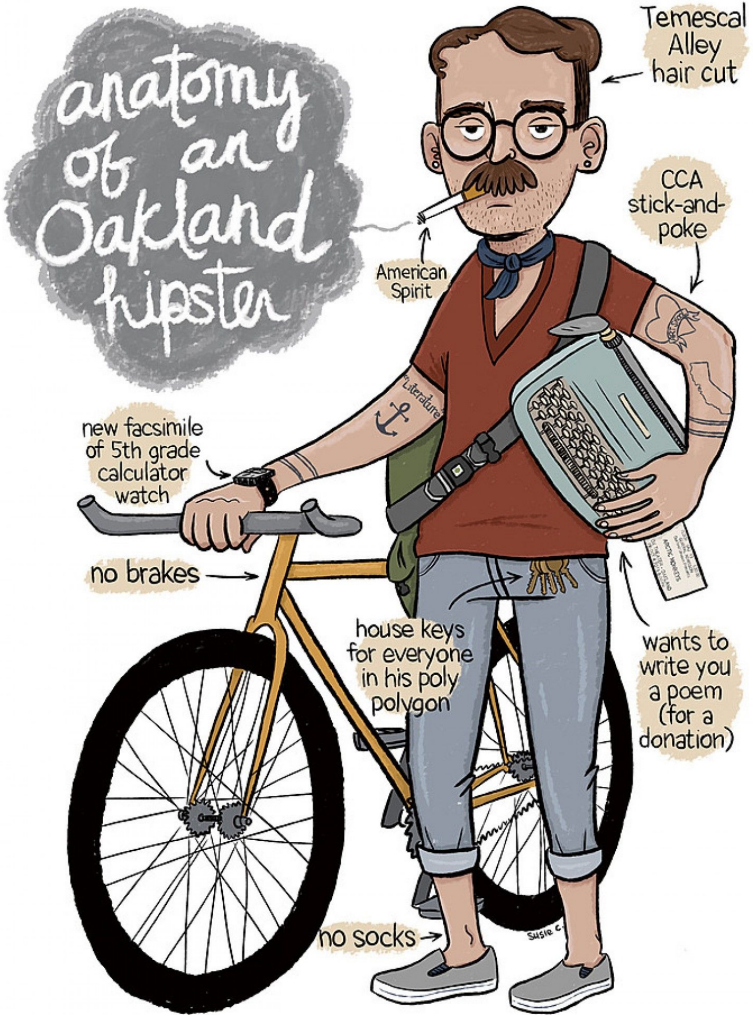
REJECTED.

“He was rejected, because he is a hipster.”

(As in “Being a hipster *caused* him to be rejected.”)

Manager’s defense: “Not at all. He just didn’t meet the dress code.”

(As in “His clothing caused him to be rejected.”)



What is a hipster?

Canadian lumberjack accidentally voted Portland's 'Hipster of the Year'

Source:
https://www.the-postillon.com/2018/04/canadian-lumberjack-hipster.html

A *hipster* is a person who diagonalizes against stereotypes.

Word History

The Original Hipsters

Hint: they carried things on their hips

There is a peculiar quotation in an issue of *The New York Tribune* from the end of 1920, which appears to be about the recent proliferation of *hipsters*:

> "How can twenty-five men keep Chicago dry, when it would take that many to watch the hipsters in one hotel dining room?" This is the question heard among those who already have obtained table reservations.
> —*New York Tribune*, 22 Dec., 1920

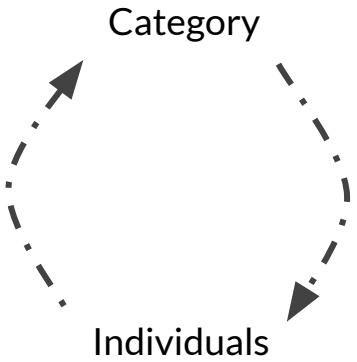
The original 'hipsters' were people who carried hip flasks during Prohibition.

Merriam-Webster SINCE 1828

Ontological instability

Taylor instability: Changing norms, theories

Hacking instability: Individuals respond to our classifications, existing categories break down as a result (“looping effect”)



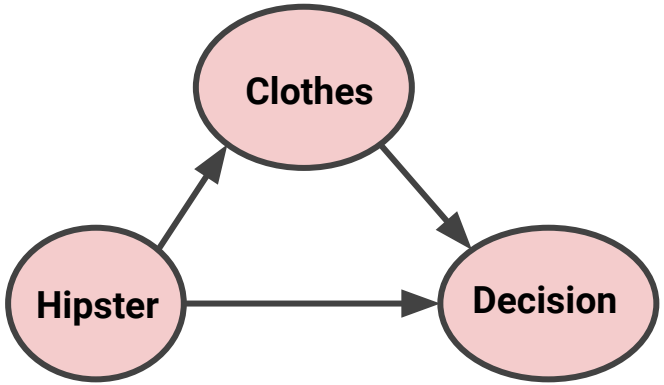
Source: Mallon. The construction of human kinds, 2019.
Hacking. Making up people, 2006.

Hypothetical ontology 1

Take 1:

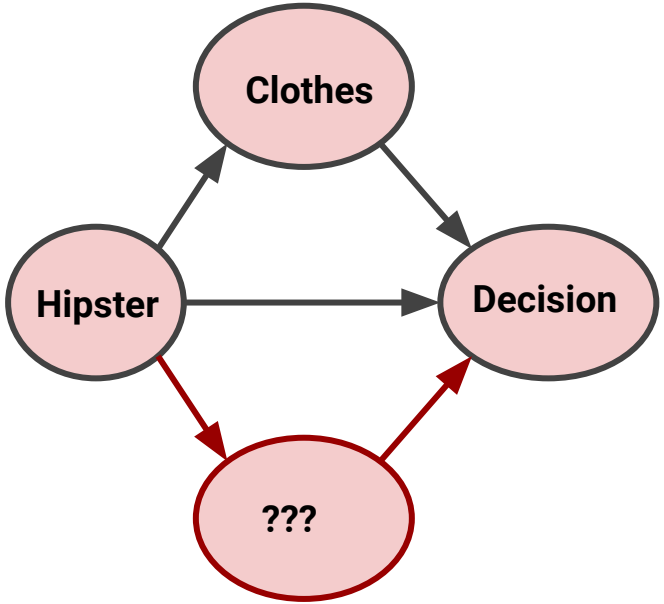
Hipster is a trait that influences various things like facial hair, clothing, place of residence, …

These variables become mediators



Hypothetical ontology 1

Puts burden of isolating right set of mediators

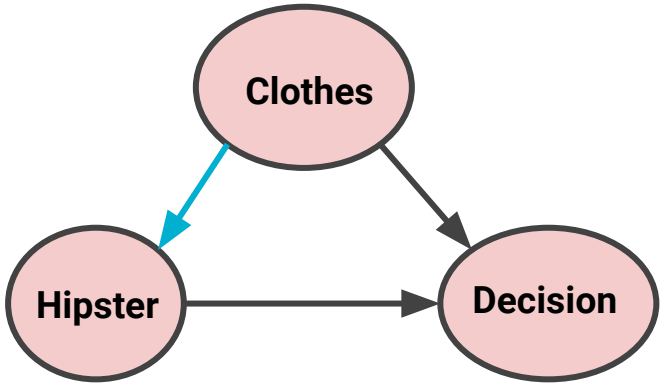


Hypothetical ontology 2

Take 2:

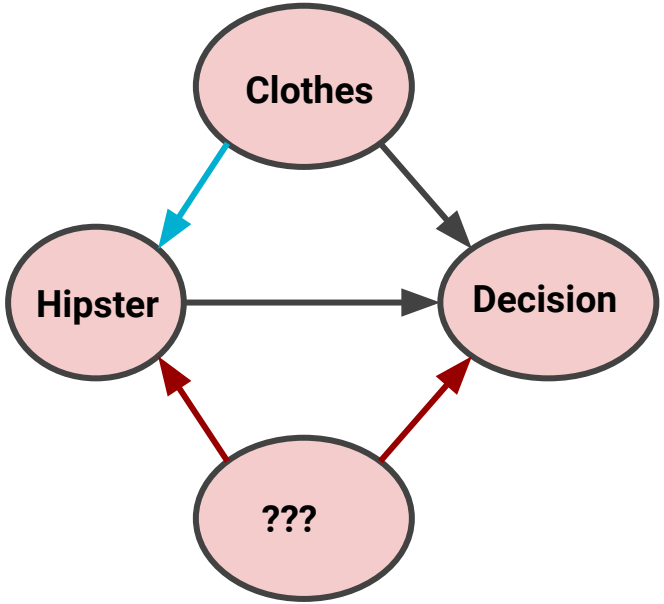
Hipster is a socially constructed from various factors, such as, beard

Various factors influence how hipster is assigned to you and how you behave in response to the categorization



Hypothetical ontology 2

Puts burden on isolating right set of confounders.



Ontology matters

Ontology 1: Clothing is a mediator

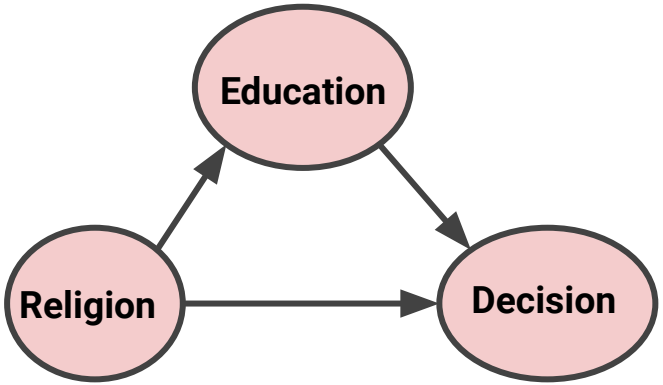
Ontology 2: Clothing is a confounder

“As you surely know by now, mistaking a mediator for a confounder is one of the deadliest sins in causal inference and may lead to the most outrageous error. The latter invites adjustment; the former forbids it.” - Book of Why, p. 276

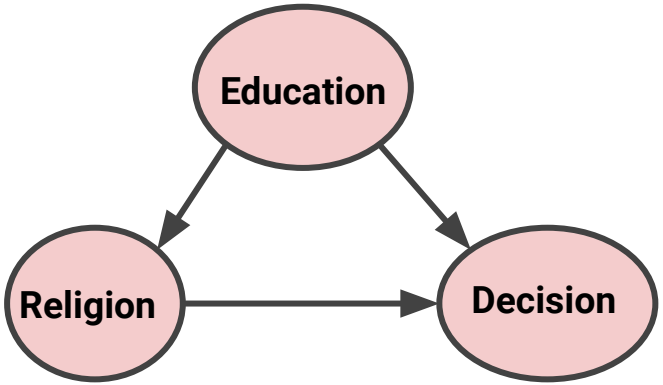
But the bigger problem is: Why can “hipster” even be a node in the first place? What does it reference? What are its “settings”?

Moving towards serious examples

“She was rejected because of her religion.”



Religion as an attribute



Proposed by Zhang-Barenboim (2018)

These competing models are manifestations of the suppressed question of what the “religion” node references in the first place.

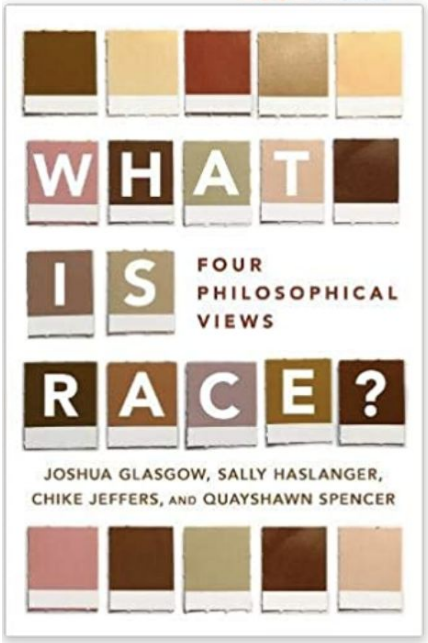
What is race?

What do we reference when we put ‘race’ in a causal model?

Race is a social construct. What race is does not follow from a biological theory.

Metaphysics: Sally Haslanger, in particular, “Ontology and Social Construction,” “Social Construction: The Debunking Project,” “Gender and Race: (What) Are They? (What) Do We Want Them To Be?,” “Oppressions,” “A Social Constructionist Analysis of Race”. Also, see work by Ron Mallon and Quayshawn Spencer.

Political Philosophy, Philosophy of Race: Tommie Shelby, Charles W. Mills, Kwame Anthony Appiah



So, where can we go from here?

The Holland (1986) position

Put as bluntly and as contentiously as possible, in this article I take the position that causes are only those things that could, in principle, be treatments in experiments.

The only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit. Statements of "causation" that involve attributes as "causes" are always statements of association between the values of an attribute and a response variable across the units in a population.

Ontological stability and manipulability

By laying out an experimental mechanism of manipulation, we hope to *certify* that the construct has a “stable enough ontology” that permits it to become a cause.

Pearl rightfully point out that manipulation is not necessary for things to become causes: “The earthquake caused the house to collapse.”

However, for social constructs it’s not clear what this kind of argument, a so-called “formation story”, would look like.

See: Hirschmann, Reed. Formation Stories and Causality in Sociology (2014).

Give up on social constructs as causes?

I’d argue this would be a serious loss.

Causal statements have political, social, and legal significance.

They can be powerful tools in locating injustice and holding people accountable.

Causal explanations are a cornerstone of the social sciences.

The Rubin-Greiner proposal

Move from race to “exposure to race” (e.g., name on a CV)

How does a decision maker respond to the “perceived race” of an individual?

Hope: Exposure to race is something we can “tweak”, “wiggle”

This is the approach taken in many audit studies and much empirical work

E.g. Bertrand, Mullainathan (2003)

It has surfaced striking evidence of racism, sexism, and discrimination

Sandra Bauer

Meryem Öztürk

Meryem Öztürk

Interview rate:	18.8%	13.5%	4.2%

Wechselbaumer (2016)

Limitations of the Rubin-Greiner proposal

Perception of the name “Meryem” on its own has no causal powers. It’s the name “Meryem” plus a whole slew of social facts that trigger a response.

But what are those facts and where do they come from? This is where we’ve pushed the metaphysical problem now.

Practically, this shows up as we try to sort out what to control for in an audit study.

Evaluates a very limited causal effect: Direct effect of a name in a world where Meryem and Sandra have identical CVs.

Cf. personal communication with Lily Hu and Issa Kohler-Hausmann.

The Hu-Kohler-Hausmann program

“Causal inference paradigms formalized for inquiries in the natural sciences (especially biological sciences) are directly transferred over to the social sciences

But social causes like race and gender are unlike more ‘standard’ causes, thereby challenging the usefulness of these theories and methods

Social categories produce causal effects because of how they are constituted by a matrix of social facts. Social constitution and causation must be united in order to understand answer social causal inference questions” - Lily Hu

Mostly work in progress, but see: 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

Conclusions

Causality clarifies issues such as confounding and mediation, but it does not resolve the normative question of what is *fair*

Put bluntly, there is no causal fairness definition either.

Causal modeling involving social constructs suffers from unacknowledged ontological and epistemological problems that are deep and challenging

Work on causality and fair decision making cannot move forward without addressing these problems.

Dynamic modeling and fairness

Recall

Statistical fairness criteria are typically thought of as near-term interventions, i.e., algorithmic constraints

As such, they’ve been criticized for ignoring feedback effects, long-term effects, and the surrounding social system

Delayed impact of fair machine learning

Do existing fairness criteria promote long-term welfare improvements for the groups they aim to protect?

Basic setup: Score R, decision rule $D = \mathbf{1}\{R > t\}$, outcome Y

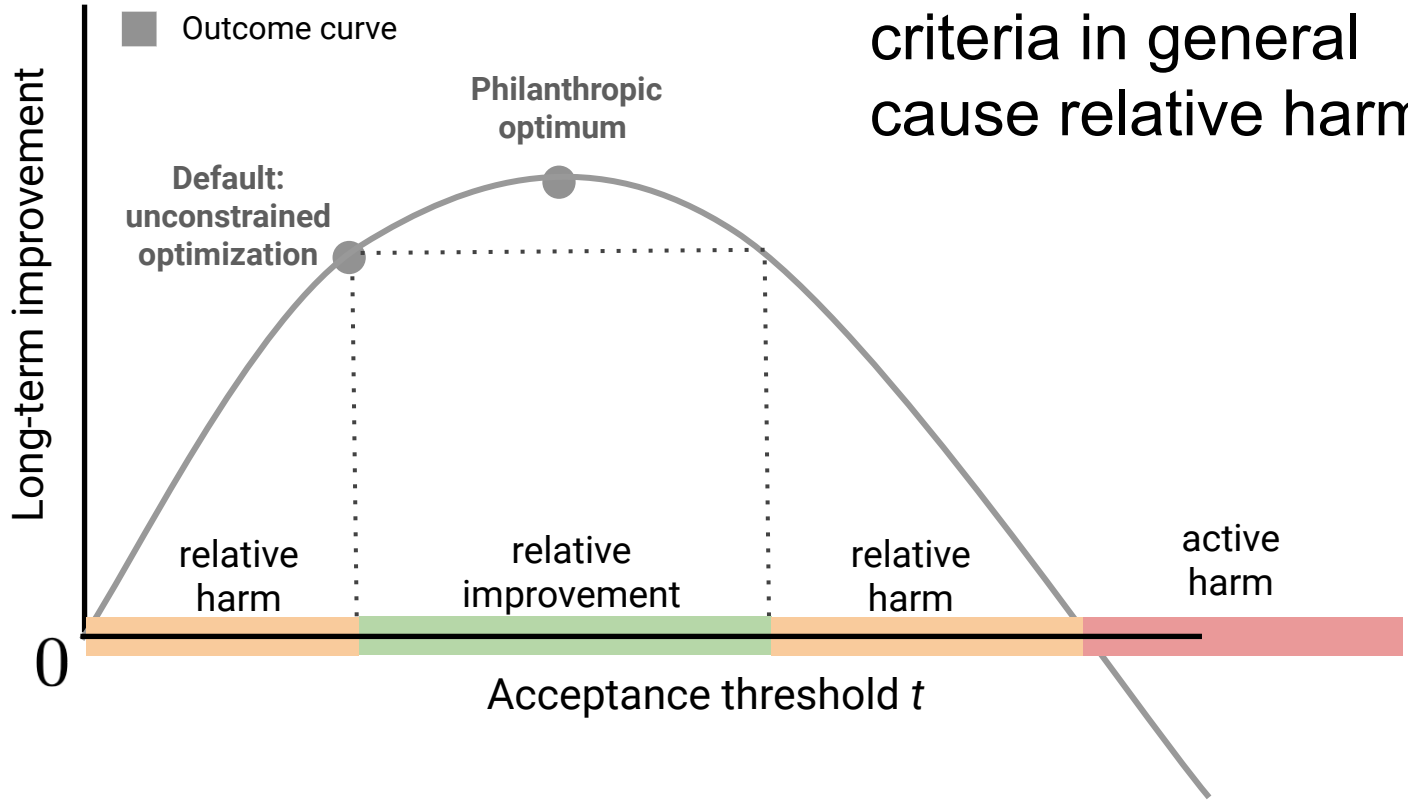
$D = 1$, $Y = 1$ improves score

$D = 1$, $Y = 0$ diminishes score

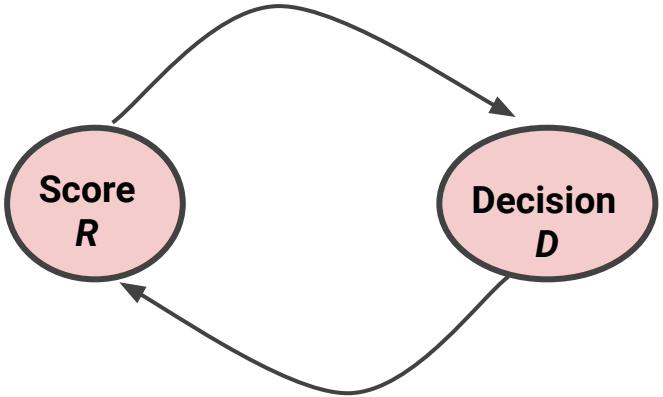
$D = 0$ leaves score the same

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

Existing fairness criteria in general cause relative harm.

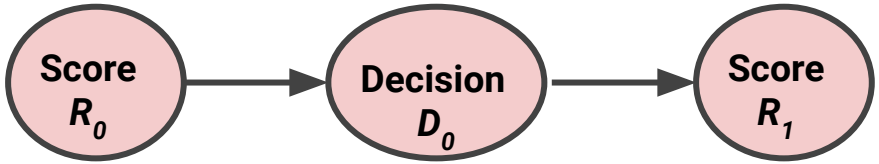


A simple dynamic behind delayed impact



Two variable dynamic model

Single time step



Unrolled acyclic graph

What about more steps?

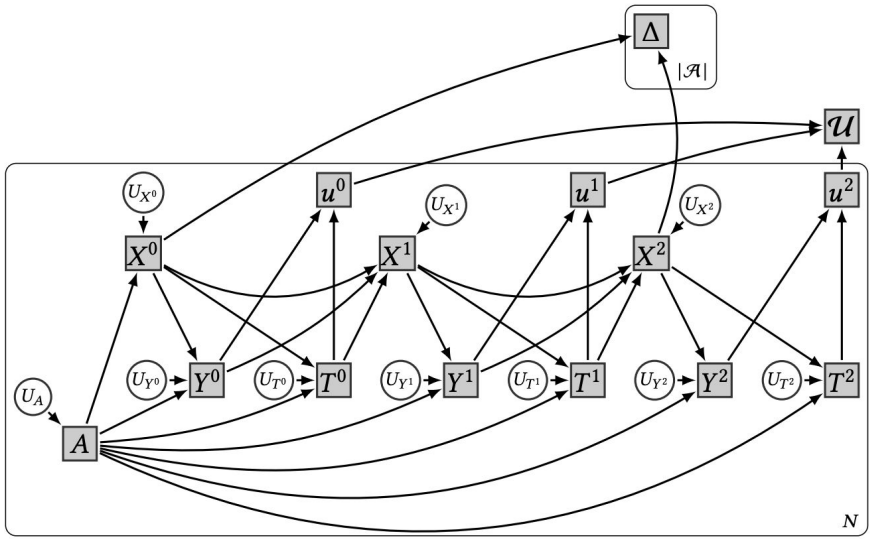


Figure 7: Phrasing the model from Liu et al. [39] as an SCM enables a multi-step extension for measuring long-term impacts, e.g., in the two-step version shown here.

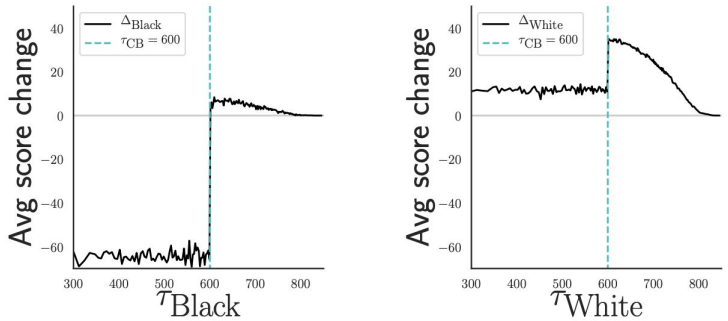
Google AI Blog

The latest news from Google AI

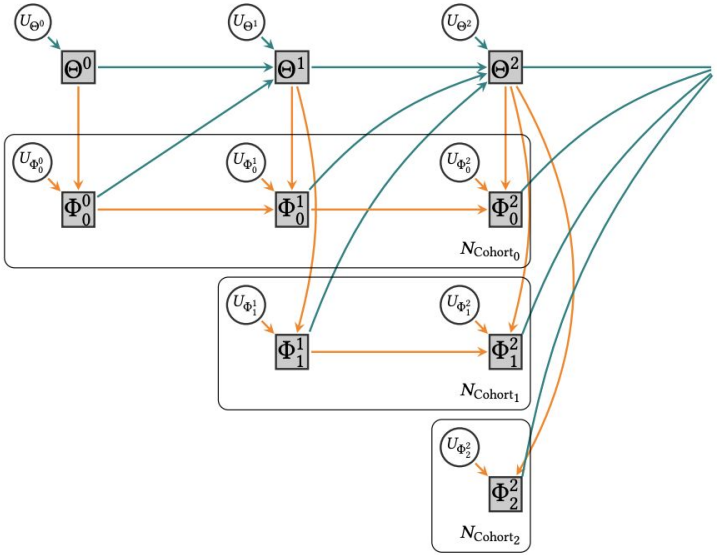
ML-fairness-gym: A Tool for Exploring Long-Term Impacts of Machine Learning Systems

Wednesday, February 5, 2020

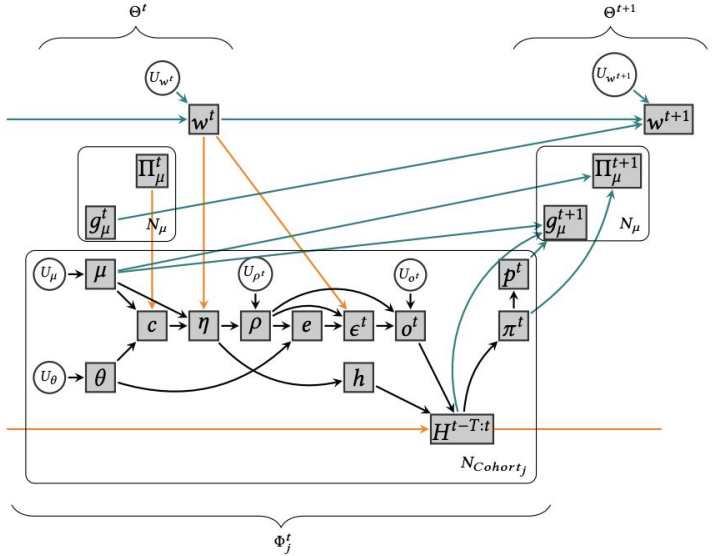
Posted by Hansa Srinivasan, Software Engineer, Google Research



Source: [Craeger, Madras, Pitassi, Zemel. Causal Modeling for Fairness in Dynamical Systems (2019)]



(a) Macro-level DAG showing how market state Θ^t and worker cohorts Φ_j^t dynamically affect one another



(b) Micro-level DAG isolating how market state Θ^t affects investment and effort levels of a single worker cohort, and how worker choices affect market state at the next step

Figure 6: SCM for the hiring model from Hu and Chen [24]. 6a shows macro-level causal assumptions. At step t the global state Θ^t of the PLM affects the choices of all cohorts of workers (a cohort denotes workers that enter the market at the same step) via wage signals (6b). The choices of investment and effort and resulting outcomes in turn affect the workers themselves in terms of hiring decisions, and the global state of the market in terms of average group reputation and performance per group. Teal arrows denote structural functions going into the global state. Orange arrows denote structural functions going into the cohort state. Black arrows denote structural functions within the cohort state. See Table 3 in Appendix B for explanation of all symbols, and Section 5 for description of the dynamics.

Source: Craeger et al. (2019)

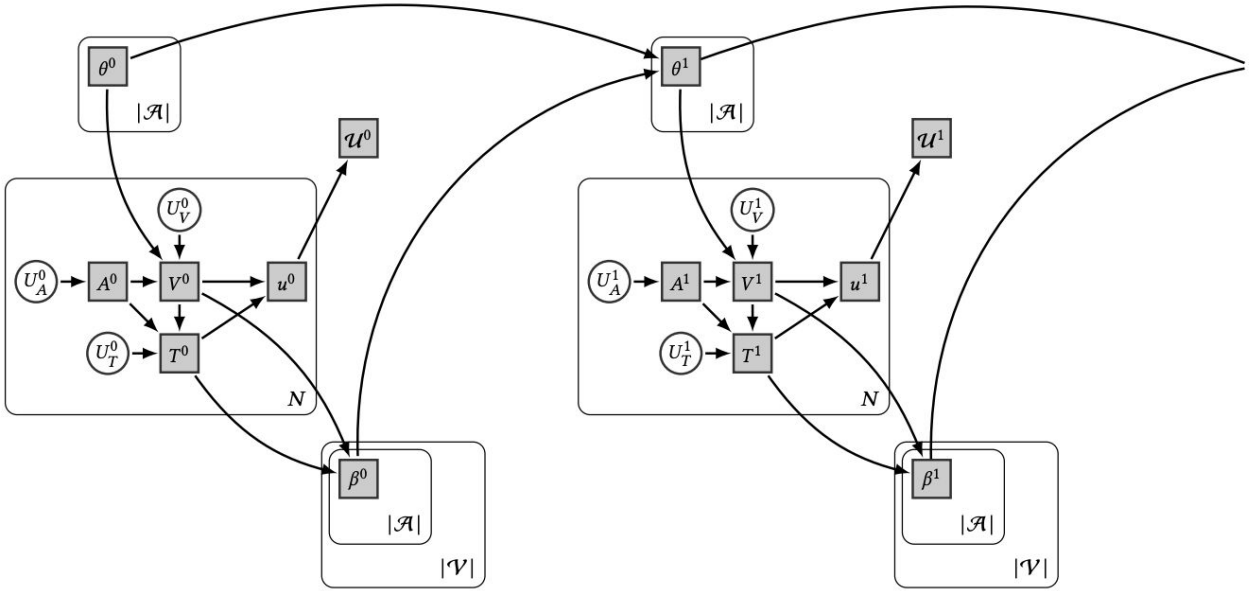


Figure 12: SCM for the group dynamics model proposed by Mouzannar et al. [46]. See Table 4 for a description of each symbol.

Source: Craeger et al. (2019)

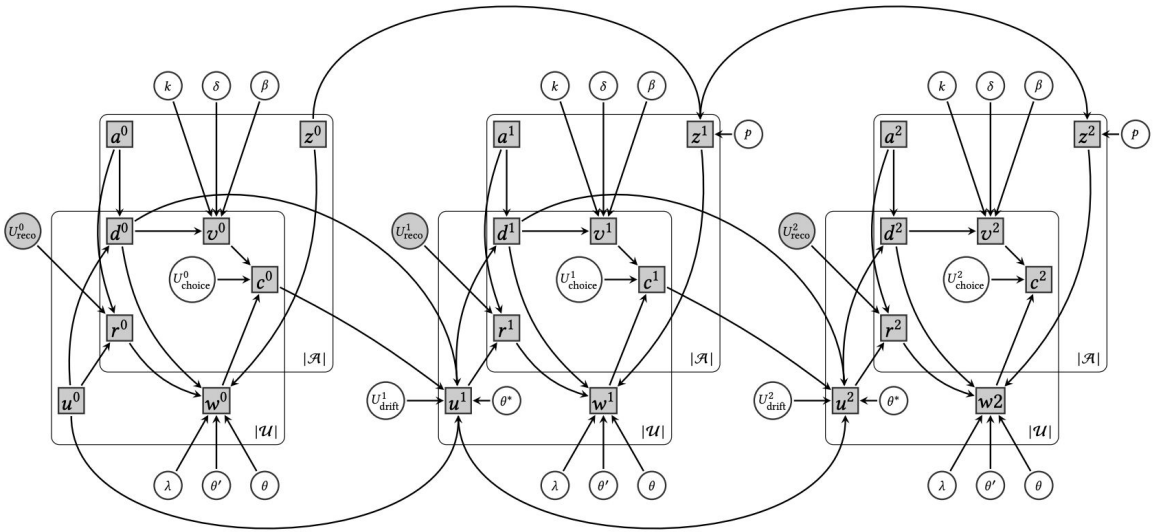


Figure 13: SCM for the news recommendation simulator model proposed by Bountouridis et al. [3]. The key dynamic modeling is in the user vectors in topic space, which drift over time towards the articles that are consumed (these in turn partially depend on the recommendations). Articles are also modeled as decaying in popularity in time. See Table 5 for explanation of all symbols.

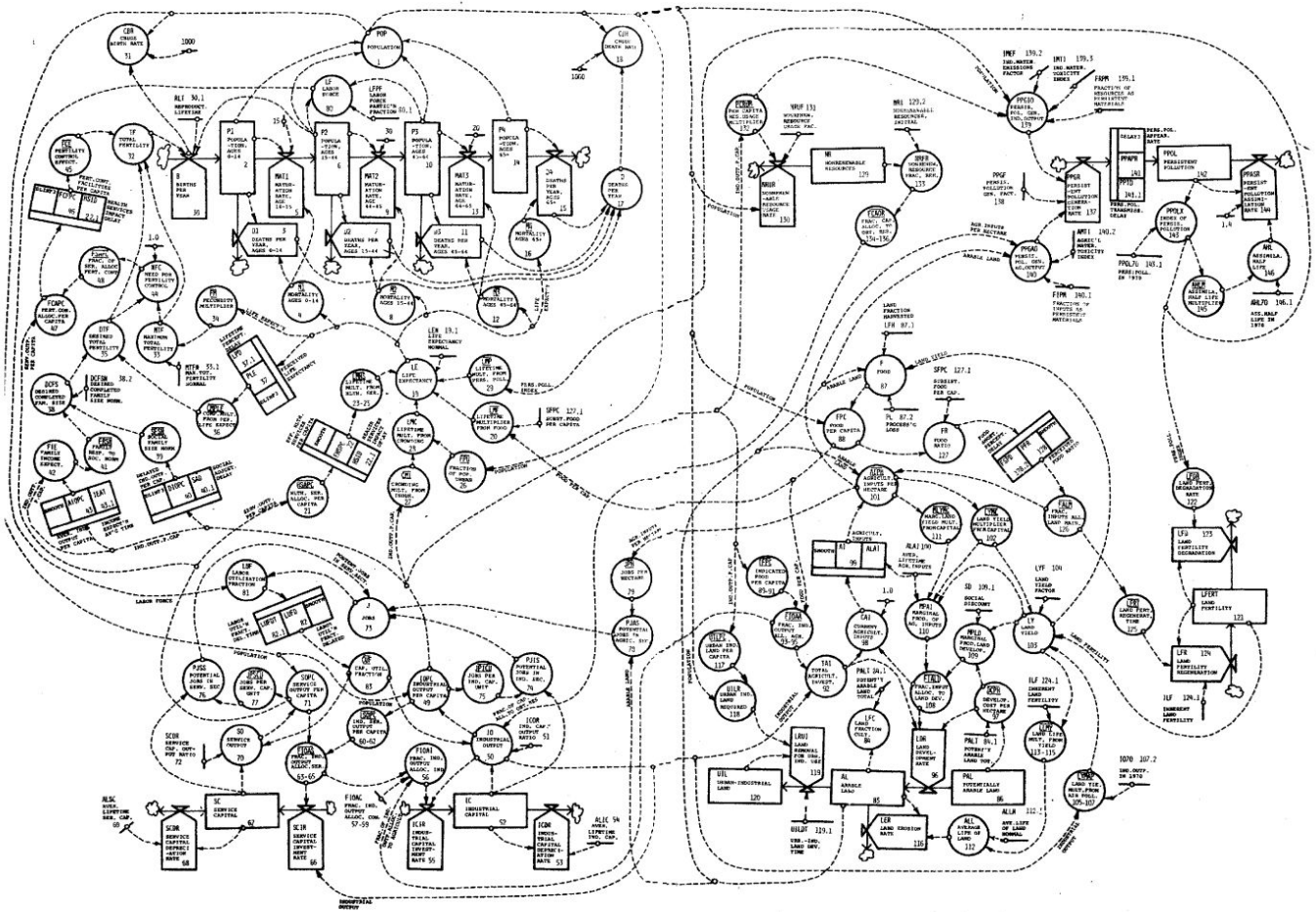
Source: Craeger et al. (2019)

OMG. This is so cool!

What if we could create dynamical models of entire sociotechnical systems, of companies, of cities, or even the whole world?

What if we use these models for policy and interventions in social systems?

It turns out, we tried that before.



"Urban dynamics"
by Jay Forrester

Late 60s

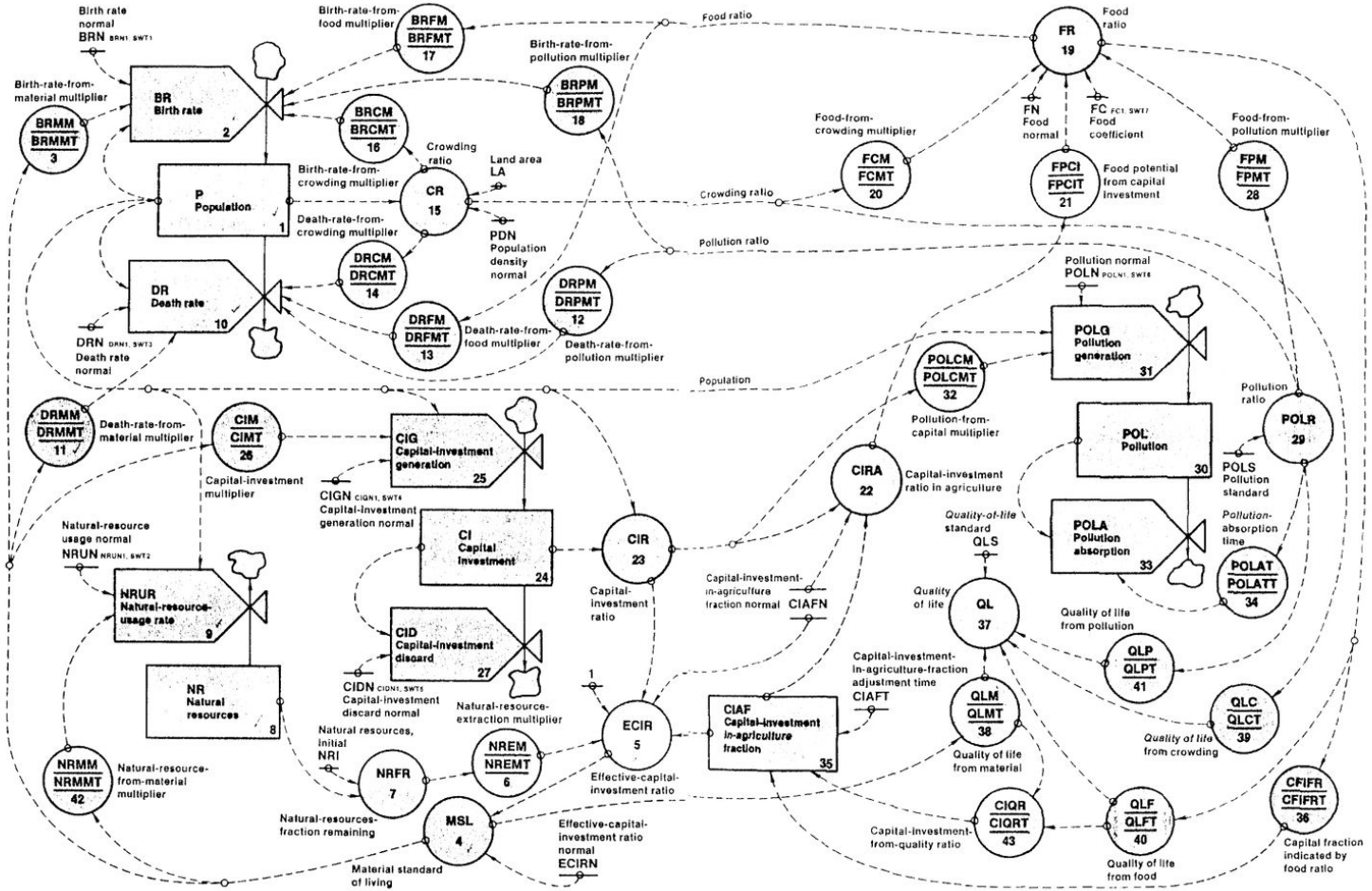
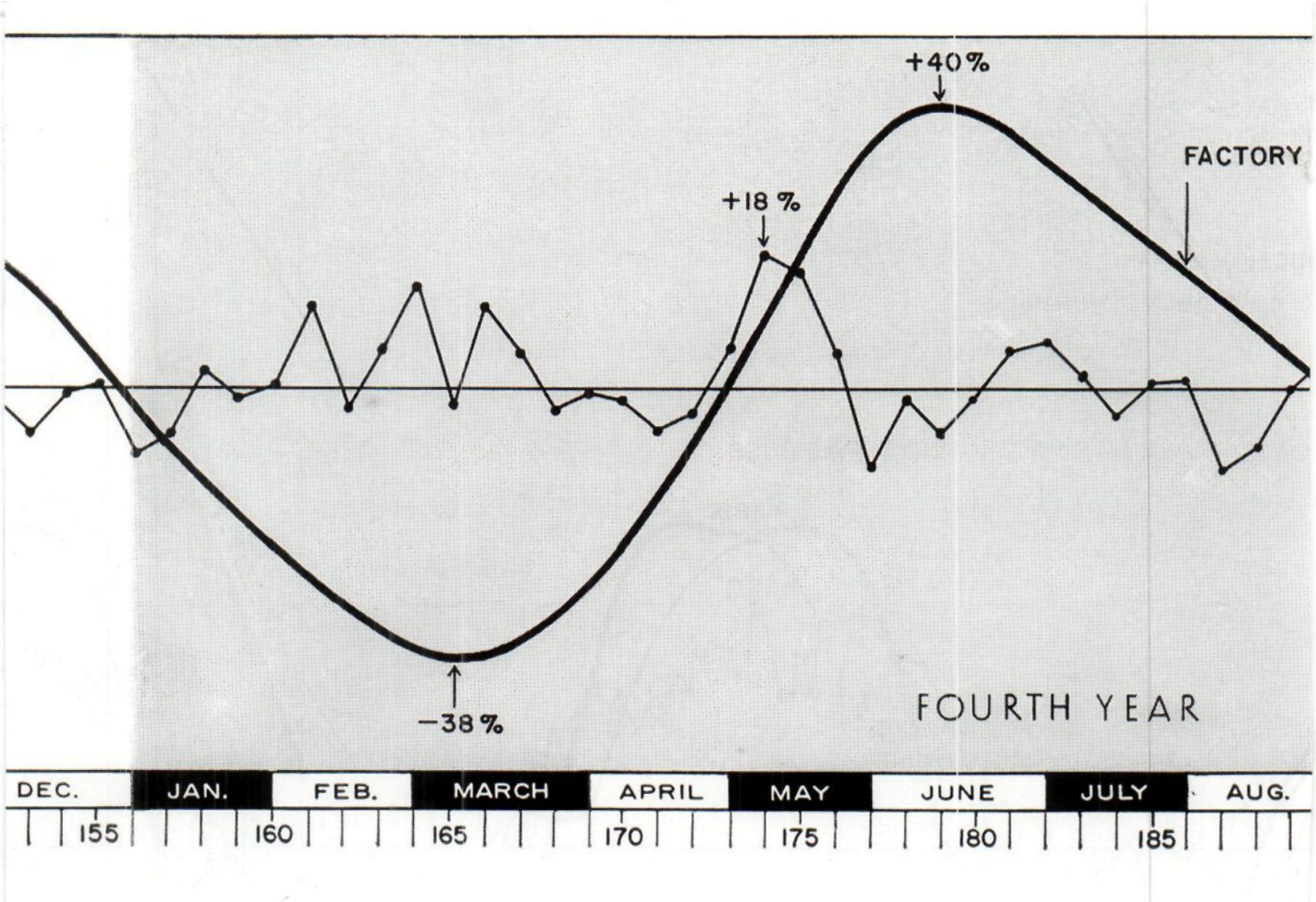


Figure 2-1 Complete diagram of the world model interrelating the five level variables — population, natural resources, capital investment, capital-investment-in-agriculture fraction, and pollution.

"World model"
by Jay Forrester

Early 70s



"Industrial dynamics"
simulation run
by Jay Forrester

1958

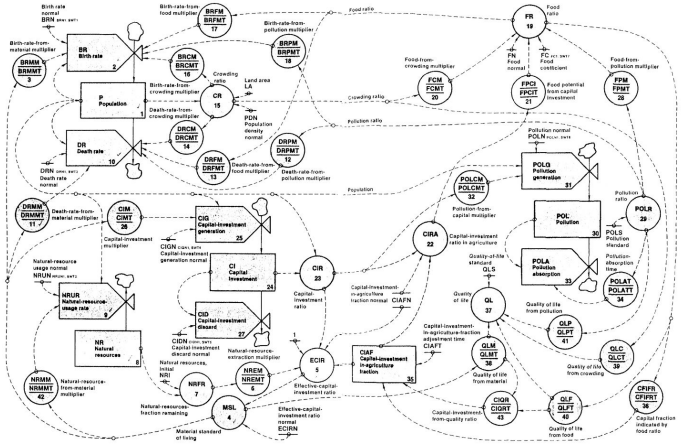
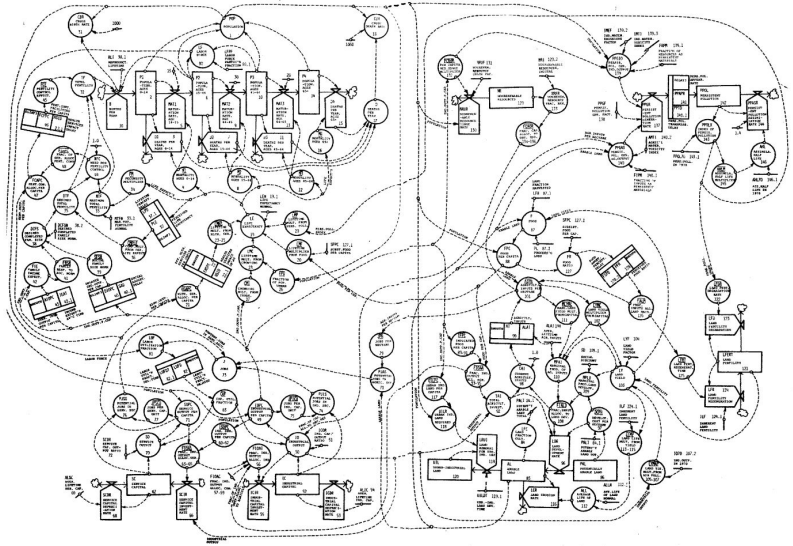
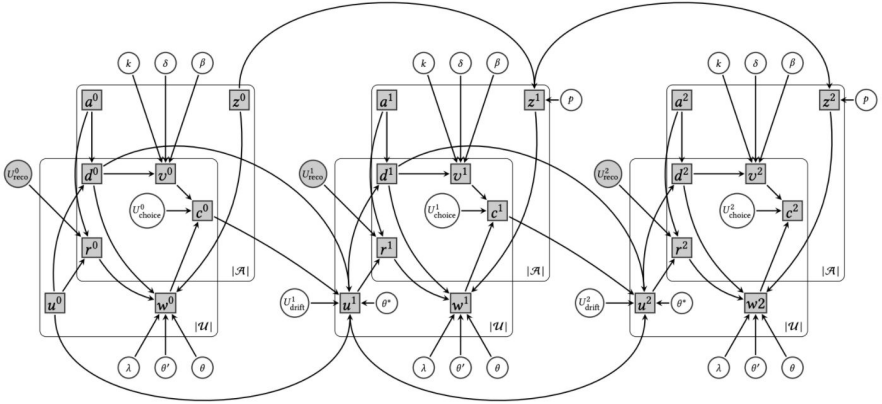
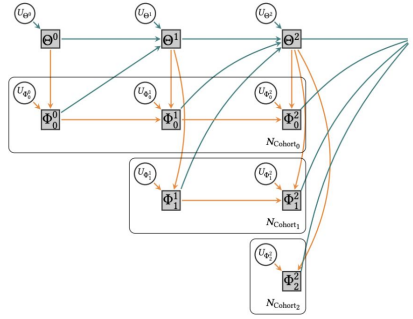
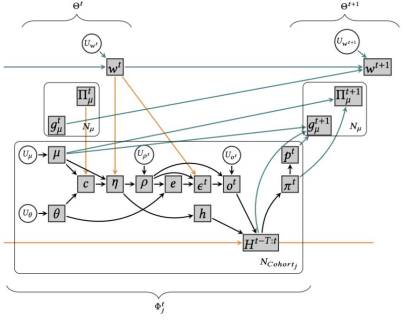


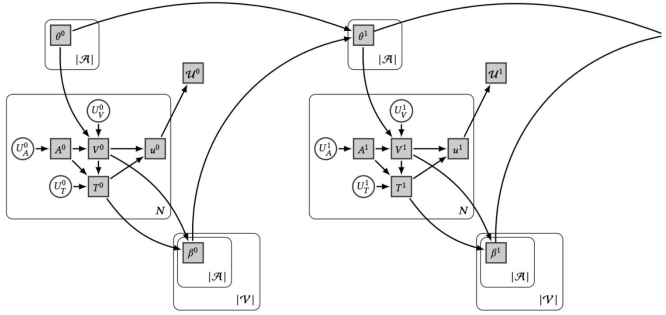
Figure 2-1 Complete diagram of the world model interrelating the five level-variables — population, natural resources, capital investment, capital investment-in-agriculture fraction, and pollution.

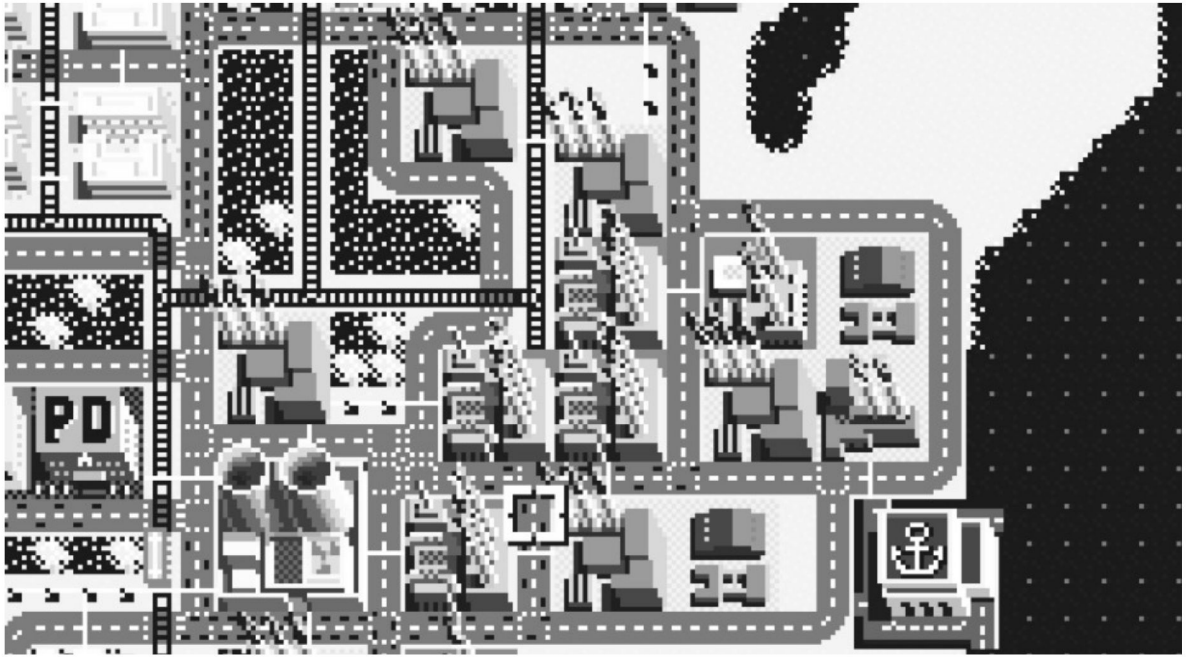












Model Metropolis

Kevin T. Baker

Behind one of the most iconic computer games of all time is a theory of how cities die—one that has proven dangerously influential.

The rebirth of system dynamics

Hypothesis:

As scholars attempt to formalize sociotechnical systems today, they create dynamical systems diagrams reminiscent of the system dynamics era.

With it comes the risk of repeating the failures of system dynamics.

But it’s not too late to learn from decades of history.

Ongoing work with Kevin T. Baker.

Failure points of system dynamics

Validation: SD models incorporated notoriously unsupported assumptions (Forrester called it “modeling the symptoms”). They also share the metaphysical problem.

Aggregation: Overly homogenous population-level models

Two Trick Pony: SD models typically exhibit only two different modes (overshoot and decline)

Pseudo-legibility: SD models purported to be legible to policy makers

How can we anticipate and mitigate these failure points?

WHYNOT

New Python package at github.com/zykls/whynot

Experimental sandbox for decision making and causal inference in dynamic environments

Intended to illustrate and anticipate failure points, test robustness, rather than for creating forecasts and policy.

Lead designer and developer: John Miller
Contributors:
Chloe Hsu, Jordan Troutman Juan Perdomo Tijana Zrnic, Lydia Liu, Yu Sun, Ludwig Schmidt

Conclusions

The critique of being too narrow motivated researchers (myself included) to embrace a socio-technical systems perspective

Sociotechnical systems present challenging territory for formal methods

They inherit the metaphysical challenges of causal model and combine it with new challenges specific to dynamical systems and control theory.

History repeats itself.

Thank you