Fairness in Dynamical Systems

Richard Zemel

CSC 2541

October 29, 2019

Overview

- Motivation
    - Fairness beyond classification: decision making & causal models
    - Long-term fairness
- Current work in fair dynamical systems
- Background: Causal DAGs
    - Interventions
    - Counterfactuals
    - Example
    - Upsides of causal DAGs
- Existing papers as causal DAGs, with policy interventions

Motivation: Fairness beyond classification

- For applications with societal impacts, data-driven prediction ***changes the environment***
    - Contrast: image classification where predictions have no effect on input images
    - Example: Lending -- Loan applicant features -> Predicted credit-worthiness -> loan approval/denial -> financial outcomes for applicant
- Not fair classification but fair ***decision making*** [Barabas et al 2018].
- Decision-making modeling captures previous fairness concerns (disparate treatment vs impact, statistical independences) but also ***causal effects*** and ***long-term outcomes***

Motivation: Long-term fairness

Automated decisions have lasting impacts

Feedback loops: many deployed ML systems make several decisions over time

Past predictions affect future state, predictions

Current fair classifiers could have long-term effects that are distinct from their short-term effects

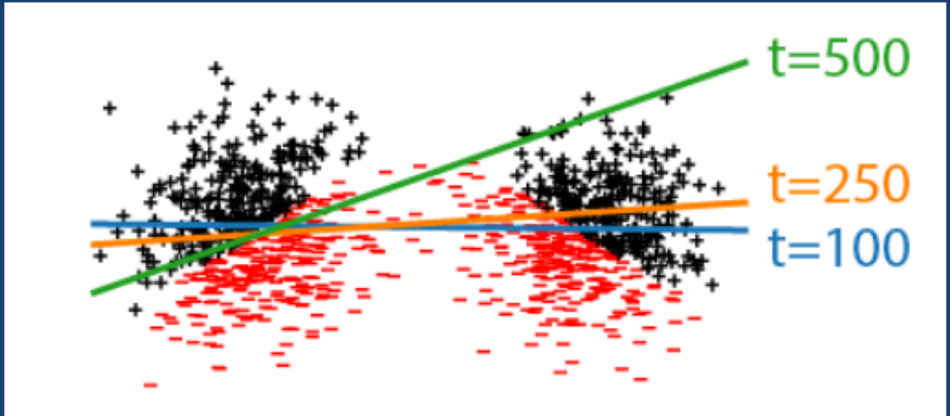
Need to take into account for long-term fair policy (“policy” := data-driven prediction)

“Fairness Without Demographics in Repeated Loss Minimization” (Hashimoto et al, ICML 2018)

- Domain: recommender systems (speech recognition, text auto-complete)
- Suppose we have a majority group ($A = 1$) and minority group ($A = 0$) – each with proportion α and unique input/output distribution
- Binary classifier repeatedly trained – w/o knowledge of group membership
- Our recommender system may have high overall accuracy but low accuracy on the minority group
- This can happen due to empirical risk minimization (ERM)
- Can also be due to repeated decision-making

Repeated Loss Minimization

- When we give bad recommendations, people leave our system
- Assume:
  - People decide to leave system independently, based on per-group expected loss
  - Classifier is not aware of group membership
- Over time, the low-accuracy group will shrink – disparity amplification



Distributionally Robust Optimization

- Upweight examples with high loss in order to improve the worst case group loss

- In the long run, this will prevent clusters from being underserved

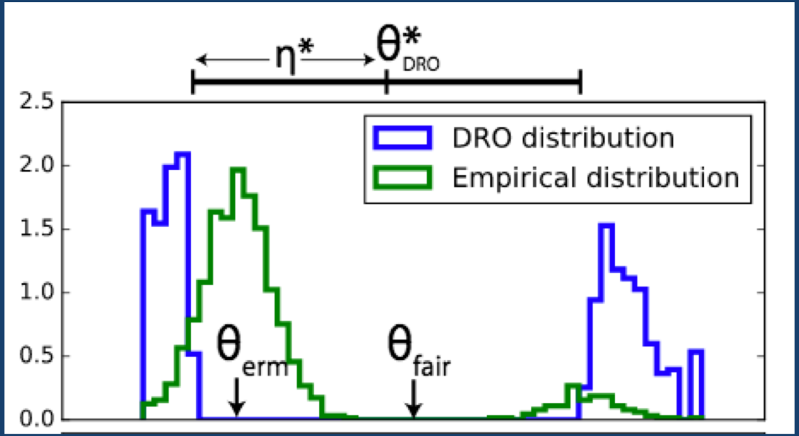
$$\mathcal{R}_{\text{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)].$$

- This ends up being equal to

$$\inf_{\eta \in \mathbb{R}} \left\{ F(\theta; \eta) := C \left(\mathbb{E}_P \left[[\ell(\theta, Z) - \eta]_+^2 \right] \right)^{\frac{1}{2}} + \eta \right\}$$

Distributionally Robust Optimization

- Upweight examples with high loss in order to improve the worst case
- In the long run, this will prevent clusters from being underserved



"Delayed Impact of Fair Machine Learning" (Liu et al, ICML 2018)

- Aim to consider feedback loops, downstream effect of decisions
- Analysis limited to single step of dynamics
- Motivating example: credit scoring
- Individual with group membership A receives credit score X, applies to bank for loan
- Bank makes binary decision T
- Binary potential outcome Y (non-default); only applies if T=1
- Loan defaults impacts bank profit, also group welfare (credit score)

Single step effects

- Loan defaults impacts bank profit, also group welfare (credit score)
- Bank makes decision based on comparing score to group-specific threshold
- Assume $\rho(x)$ is the probability of non-default for score x
- Expected utility to bank depends on $u_{+/-}$ (profit/loss based on repay/default)

$$\boldsymbol{u}(x) = u_+\boldsymbol{\rho}(x) + u_-(1 - \boldsymbol{\rho}(x))$$

- Score change model similar, depends on credit score change with repay/not

$$\boldsymbol{\Delta}(x) = c_+\boldsymbol{\rho}(x) + c_-(1 - \boldsymbol{\rho}(x))$$

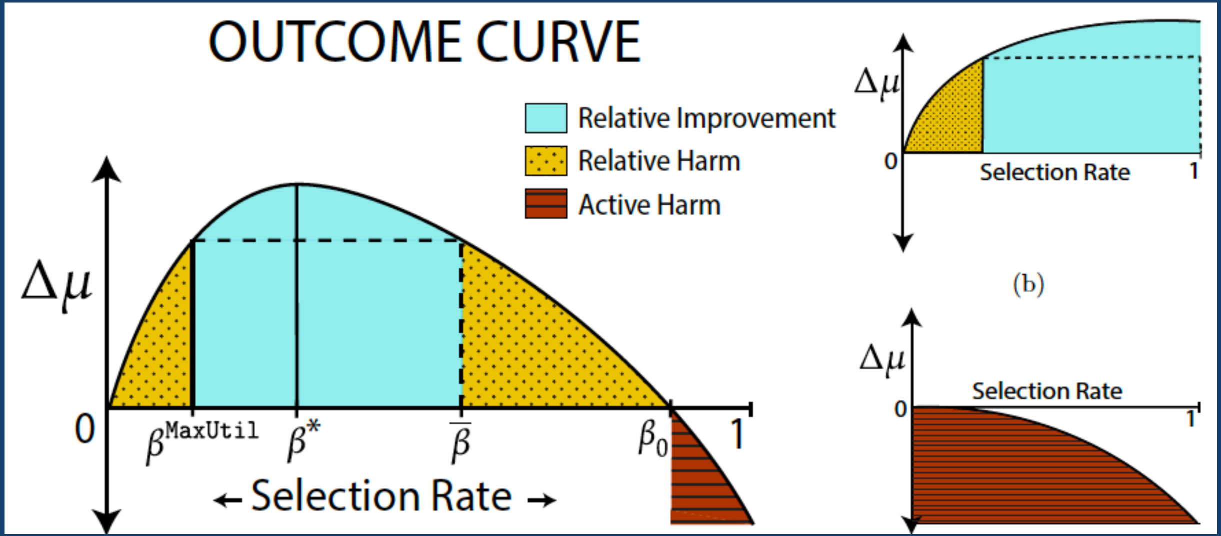
- Different thresholds satisfy different criteria: maximizing profit; demographic parity; equal opportunity

Policy impact on group

- Key statistic – change in mean score for group.

$$\Delta\mu_j(\tau) := \sum_{x\in\mathcal{X}} \pi_j(x)\tau_j(x)\Delta(x)$$

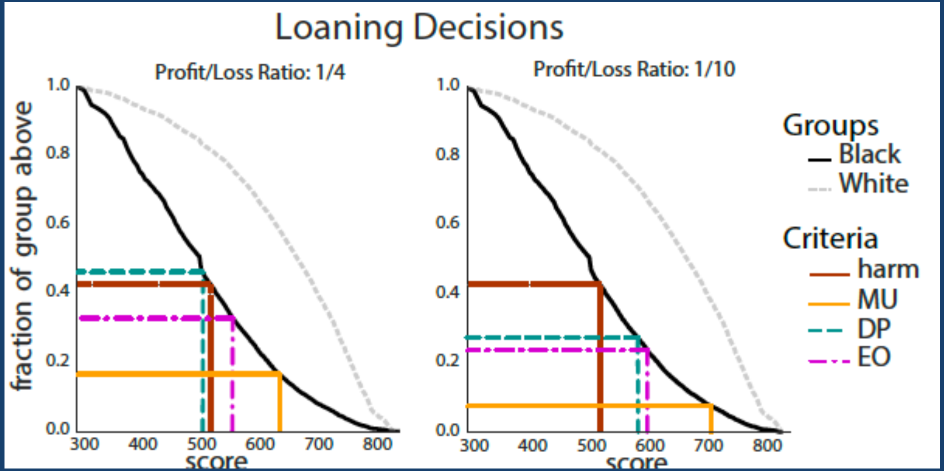
- Compare outcome for group relative to utility maximizing policy

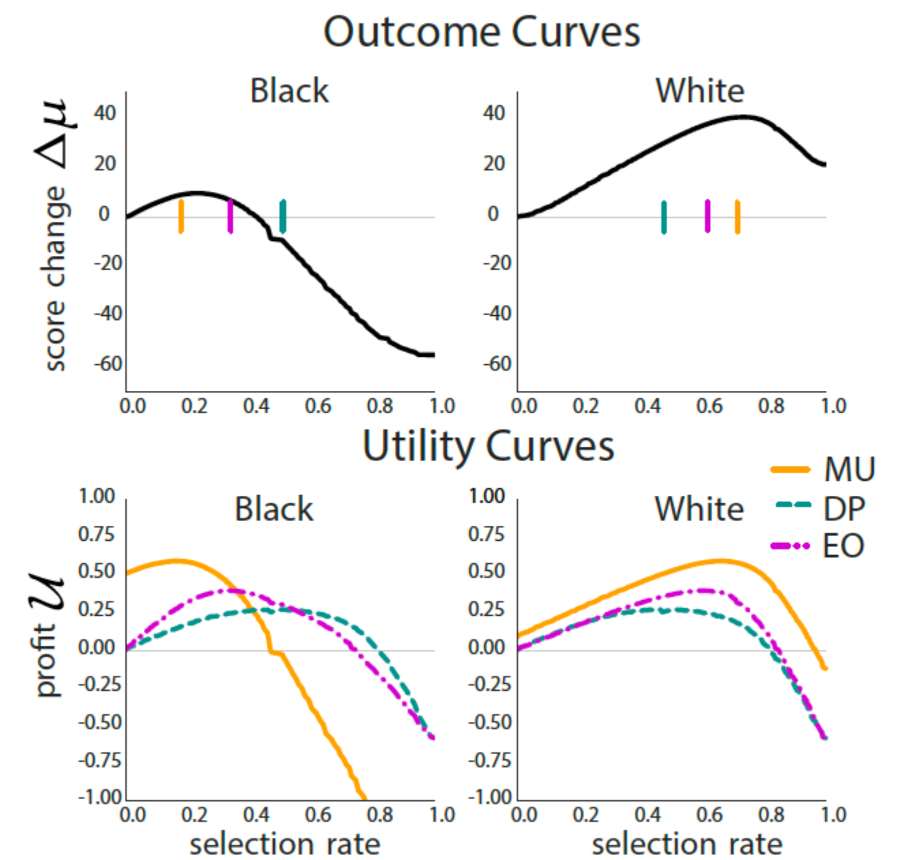


Simulation

- FICO score data – similar repay prob per group, different score histograms

- Set parameters (such as c_+ / c_-)



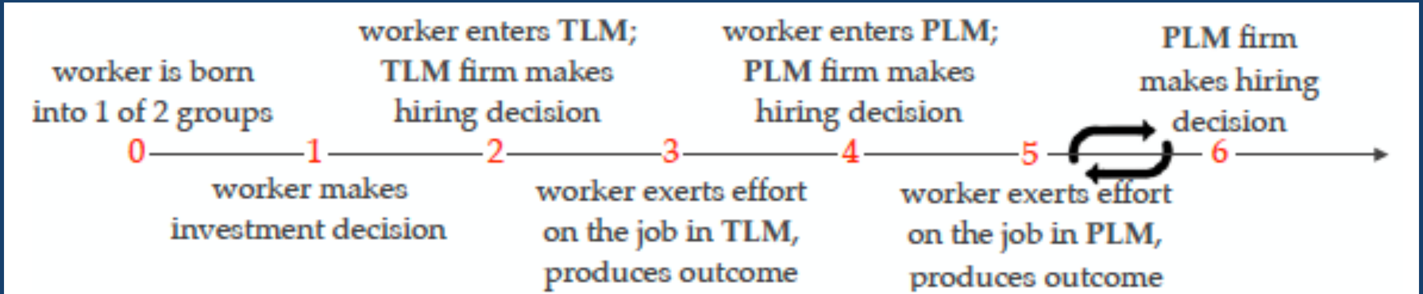


"A Short-term Intervention for Long-term Fairness in the Labor Market" (Hu & Chen, WWW 2018)

- Addressing racial inequities in labor market
- Dynamic reputational model – reinforcing nature of asymmetric outcomes, based party on group's different access to resources, investment
- Cohort of workers initialized with attributes \phi, journeys thru labor markets:
  - Temporary Labor Market – ensure statistical parity of groups entering market
  - Permanent Labor Market – firms hire who they want
- Hiring markets have global state – wages, reputations and proportion of good workers in PLM per group
- Long-term aim: group equality in labor market outcomes

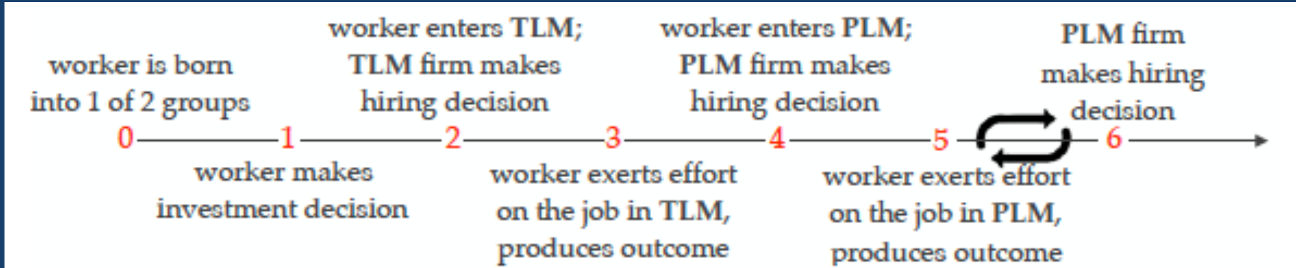
Setup

- N workers pass thru labor market over time
- Number of workers in each of 2 groups stable
- Worker abilities same, stable across groups; reputations vary – depends on proportion of workers producing good outcomes
- Workers select education investment – trade off cost versus expected reward (wages) – hired into TLM based on investment level



Setup (cont.)

- Workers born with per-group ability level \theta
- Workers from disadvantaged groups face higher costs of investment
- At each step workers respond to current wages by exerting effort (high qualified and ability workers can exert high effort w/ low cost)
- Worker's effort leads to outcomes, which are accumulated to form reputation
- Hired into PLM based on reputation -- > affects g, quality of workers→ affects wages w (more good workers will lower wages)



Hiring dynamics

Proportion of workers w/ good outcomes

$$g_t^\mu = p_H[1 - F(\widehat{\theta_Q})\gamma_t^\mu - F(\widehat{\theta_U})(1-\gamma_t^\mu)] + p_Q F(\widehat{\theta_Q})\gamma_t^\mu + p_U F(\widehat{\theta_U})(1-\gamma_t^\mu)$$

$$\text{where } \widehat{\theta_\rho} = e_\rho^{-1}(w_t(p_H - p_\rho))$$

$$\text{and } g_{t'} = \sigma_\mu \ell g_{t'}^\mu + (1-\sigma_\mu)\ell g_{t'}^\nu$$

Notation	Significance
$F(\theta)$	CDF of ability levels θ
π^μ	group μ reputation
σ_μ	group μ population share
w_t	wage at time t
g_t^μ	proportion of group μ workers producing good outcomes at time t
η	investment level
p_H, p_Q, p_U	probability of producing G given effort level
$c_{\pi_t^\mu}(\theta, \eta)$	cost of investment
$\gamma(\eta)$	probability of being qualified
$\rho \in \{Q, U\}$	hidden qualification status
$e_\rho(\theta)$	cost of effort exertion
Π_i^t	individual reputation at time t

- Argue unconstrained dynamics produce inequality
- Disadvantaged workers less likely to invest → leads to worse outcomes → lower reputation → raise investment cost
- If TLM must hire equal numbers of workers per group, will carry over to PLM

Fairness & Causality

- Many fairness problems (e.g., loans, medical diagnosis) are actually causal inference problems
- We talk about the label Y – however, this is not always observable
- For instance, we can’t know if someone would return a loan if we don’t give them one
- This means if we just train a classifier on historical data, our estimate will be biased (biased both in the fairness sense and the technical sense)

- General takeaway: if your data is generated by past decisions, think very hard about the output of your ML model
- Now we can re-examine the fair dynamics models from causal perspective

Motivation: Off-policy evaluation

Implementing “fair” policies in production is high-risk

- Bad assumption or hyperparameters could harm users

We want to know how a new (“fair”?) policy will do in production without running experiments, control trials

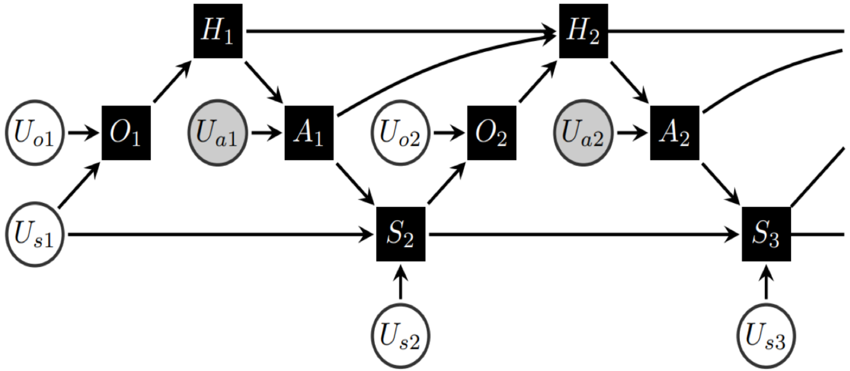
Only data from the old policy are available

Framework for dynamical fairness models

Markov decision processes (MDPs) are a natural model for sequential decision making

- Optimize policy (state -> action mapping) to maximize expected reward
  - Open research question: long-term definitions of fairness

Adopt causal formulation – one modeling framework is Structural Causal Models (SCMs)



Buesing et al 2018

Background: PGMs vs. SCMs

Probabilistic Graphical Models (PGMs) encode the conditional independences in a data generative process

- Good for ***inference problems***

Structural Causal Models (SCMs) encode conditional independencies ***and*** causal assumptions

- Structural equations specify functional form for causal mechanisms
  - Y = f_Y(U_Y, X, T)
- Good for ***intervention problems***

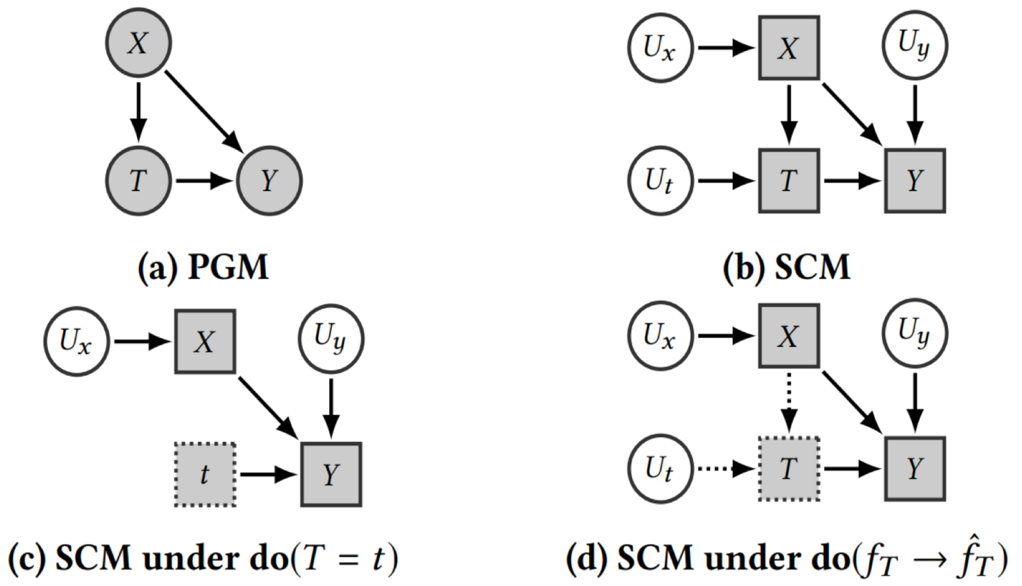


Figure 2: Treatment model expressed as PGM (2a) and SCM (2b). We also show the SCM under atomic intervention (2c) and policy intervention (2d).

Interventions

How do outcomes change in response to a forced change to the environment? (contrast against conditioning)

Atomic interventions

- change the value of one variable
- remove influence of parents

Policy interventions

- Change the functional form of one structural equation
- For example change a naive policy to a "fair" one

Multiple interventions model distinct strategic actors in the environment

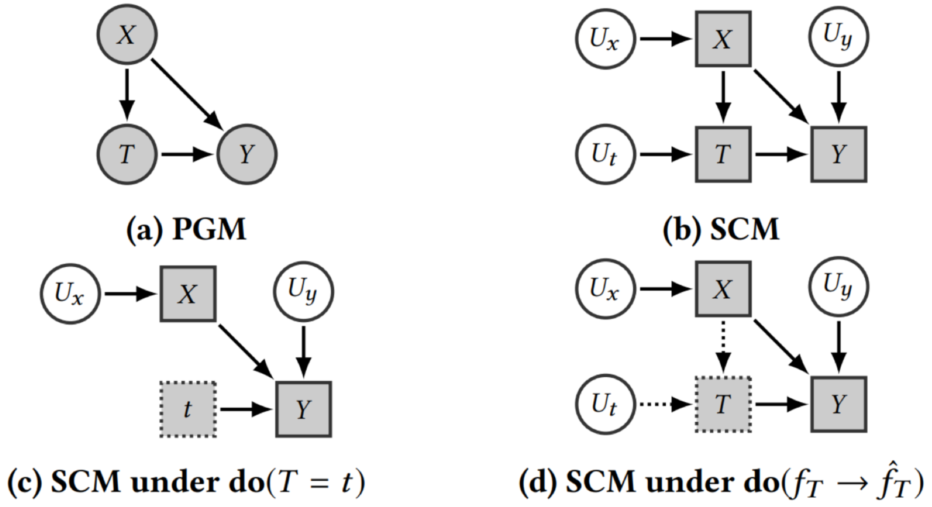


Figure 2: Treatment model expressed as PGM (2a) and SCM (2b). We also show the SCM under atomic intervention (2c) and policy intervention (2d).

Counterfactuals

Using observations to infer the scenario, how would the outcomes have been different under intervention?

1. **Infer:** Condition on observations and infer distribution over exogenous noise (i.e. latents)

2. **Intervene:** Carry out an atomic or policy intervention

3. **Outcomes**: Re-sample exogenous noise and compute outcomes

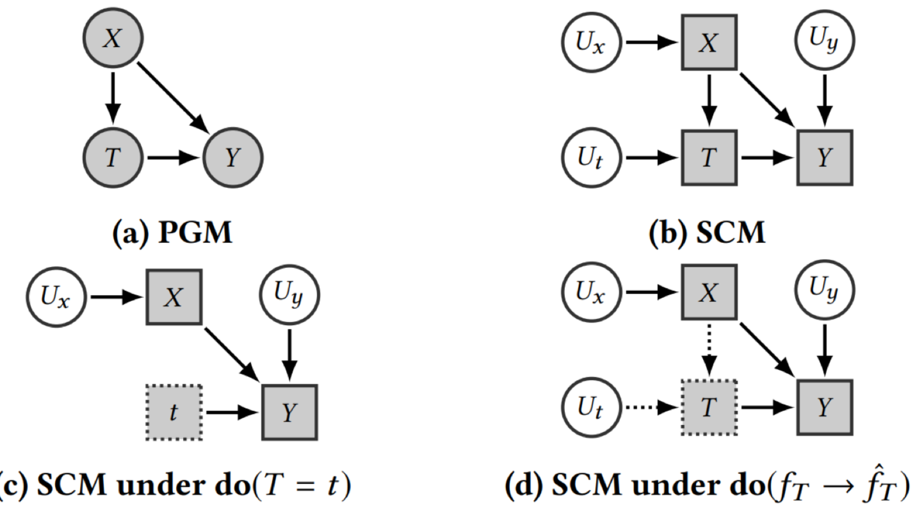


Figure 2: Treatment model expressed as PGM (2a) and SCM (2b). We also show the SCM under atomic intervention (2c) and policy intervention (2d).

Ex: Treatment model

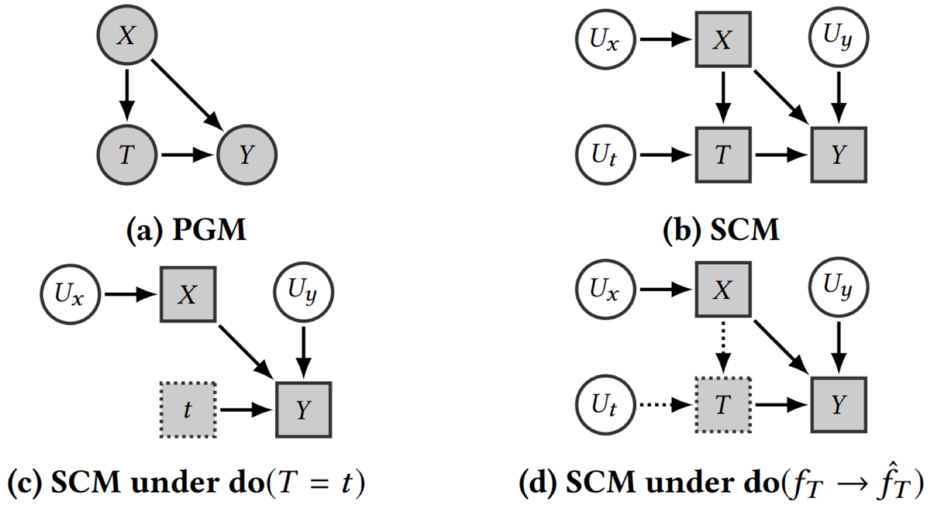


Figure 2: Treatment model expressed as PGM (2a) and SCM (2b). We also show the SCM under atomic intervention (2c) and policy intervention (2d).

X represents a confounding covariate

T represents treatment

Y represents outcome

Certain choices of ***p*** induce ***Simpson's paradox***, where ***p(Y|T)*** differs from ***p(Y|T,X=x)***

Ex: Treatment model

X represents a confounding covariate

T represents treatment

Y represents outcome

Certain choices of p induce ***Simpson's paradox***, where p(Y|T) differs from p(Y|T,X=x)

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

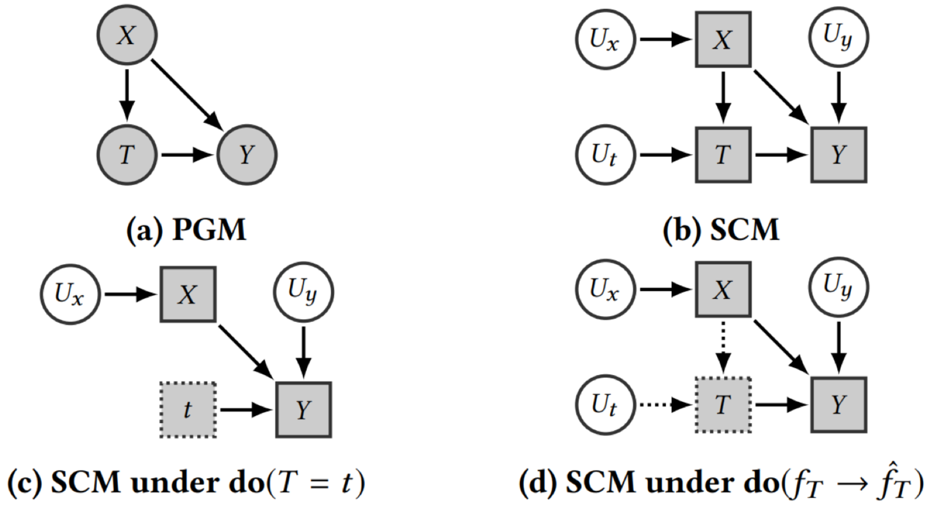


Figure 2: Treatment model expressed as PGM (2a) and SCM (2b). We also show the SCM under atomic intervention (2c) and policy intervention (2d).

Ex: Treatment model

joint:

$$p(X, T, Y) = p(X)p(T|X)p(Y|X, T)$$

conditional:

$$p(Y|T = t) = \mathbb{E}_{p(X|T=t)}\left[p(Y|X, T = t)\right]$$

interventional:

$$p_{\mathrm{do}(T \to t)}(Y|T = t) = \mathbb{E}_{p(X)}\left[p(Y|X, T = t)\right]$$

causal effect:

$$\mathbb{E}_{p_{\mathrm{do}(T \to t')}}\left[Y|T = t'\right] - \mathbb{E}_{p_{\mathrm{do}(T \to t^*)}}\left[Y|T = t^*\right]$$

counterfactual:

$$p_{\mathrm{do}(T \to t')|Y=y^*}(Y|T = t') = \mathbb{E}_{p(X|Y=y^*)}\left[p(Y|X, T = t')\right]$$

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

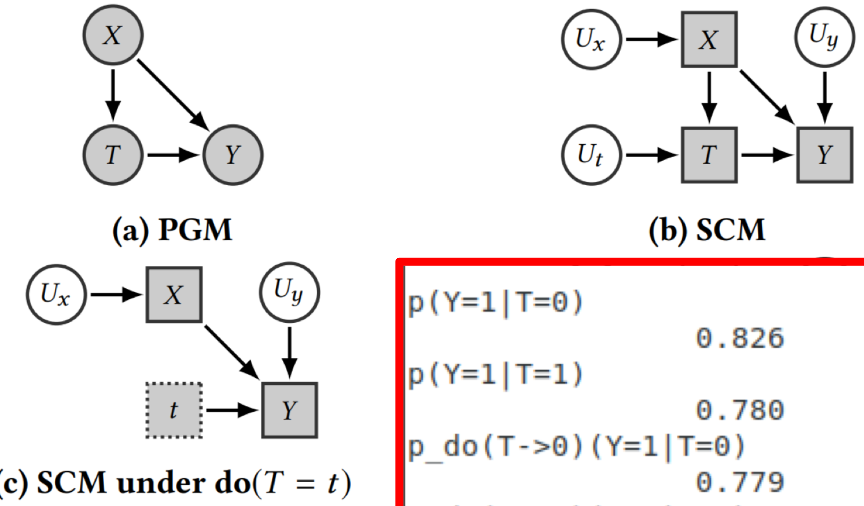


Figure 2: Treatment model e
(2b). We also show the SCM
and policy intervention (2d).

```
p(Y=1|T=0)
                    0.826
p(Y=1|T=1)
                    0.780
p_do(T->0)(Y=1|T=0)
                    0.779
p_do(T->1)(Y=1|T=1)
                    0.833
p_do(T->0|Yobs=1)(Y=1|T=0)
                    0.775
p_do(T->1|Yobs=1)(Y=1|T=1)
                    0.828
```

Fair ML: Dynamical systems and causality

Dynamical Systems

- Economics models for long-term policy effects, e.g., affirmative action [Coate and Lowry 1993, Foster and Vohra 1992]
- Feedback loops [Lum and Isaac 2017]
- Fair bandits [Joseph et al 2016] and RL [Jabbari et al 2017] algorithms
- Applications described above (and below)]
- Fairness gym: datasets -> simulation

Causality & fairness

- Fairness as counterfactual stability [Kusner et al 2017]
- Fair feature selection and adjustment given causal DAG [Kilbertus et al 2017]
- Fair inference [Nabi and Shipser 2018]

Causal Modeling in ML

Causal effect estimation

- Propensity scoring [Rosenbaum and Rubin 1983]
- Latent variable models for effect estimation [Louisoz et al 2017, Madras et al 2018]
- Measuring path-specific causal effects [Nabi and Shipster 2018]

Policy evaluation and optimization

- Refactor POMDPs as SCMs for evaluation and policy iteration via counterfactuals [Buesing et al 2018]
- Robustness of counterfactual policy evaluation to model misspecification [Oberst and Sontag 2019]

Why Causal DAGs for Fairness?

1. ***Visualization***
    a. exposes assumptions underlying the model
    b. communicates its content to others, especially non-mathematical stakeholders

2. ***Introspection***
    a. explicit causal assumptions invite ***scrutiny*** by modelers, domain experts
        i. safeguard against blind solutionism (***don't overclaim*** in fairness papers)
    b. Inspecting CDAG of existing model can suggest new policies, interventions, and robustness questions

3. ***Evaluation***
    a. Specifying a joint distribution as a causal DAG enables causal reasoning.
        i. Off-policy evaluation: estimate policy impact without incurring risk of deployment
        ii. Simulate "what-if" scenarios with counterfactual generation

Limitations of Causal DAGs

1. ***No guarantees under incorrect assumptions***
   a. Causal assumptions are often untestable (especially in fairness applications)
      i. Emphasizes dependence on a ***correct*** domain expert
   b. degrees of misspecification: graph structure mismatch vs structural equations mismatch
   c. A special concern: ***unobserved confounding***

2. ***Sophisticated models induce tangled graphs***
   a. For effective communication to non-experts we need the right level of abstraction
   b. Inspecting CDAG of existing model can suggest new policies, interventions, and robustness questions

3. ***Lack of tooling***
   a. Need flexible inference/intervention/simulation for counterfactual reasoning

Causal DAG Formulations of Existing Work

Domain	**Paper**	**Features**
Lending	Liu et al 2018. *Delayed Impact of Fair Machine Learning.*	* Dynamics in individual credit scores * Treat bank policy (loan predictor) as supervised problem * Evaluated one-step fairness of various constrained classifiers
Repeated classification	Hashimoto et al 2018. *Fairness without demographics in repeated loss minimization.*	* Demographic group mixture model * Group membership unobserved * Dynamics in group sizes * Evaluated learning via distributionally robust optimization
Hiring	Hu and Chen 2018. *A short-term intervention for long-term fairness in the labor market.*	* Models strategy of employees & employers * Hiring model with temporary and permanent workers * Evaluated effectiveness of intervention in short-term market

Liu et al 2018
Delayed Impact of Fair ML

Dynamics in individual credit scores

Treat bank policy (loan predictor) as supervised problem

Evaluated one-step fairness of various constrained classifiers

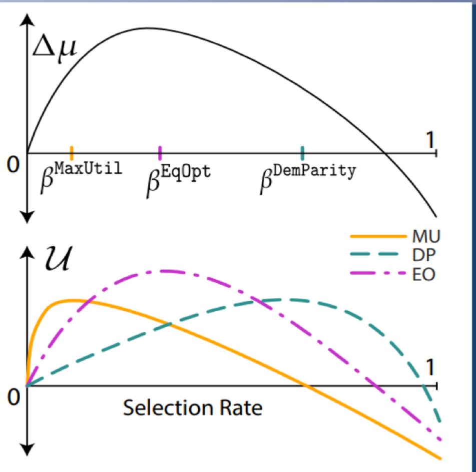
Structural eqns:

Bank policy $T = f_T(U_T, A, X)$

Potential outcome $Y = f_Y(U_Y, X, A)$

Next-step score $\tilde X = f_{\tilde X}(Y, T, X)$

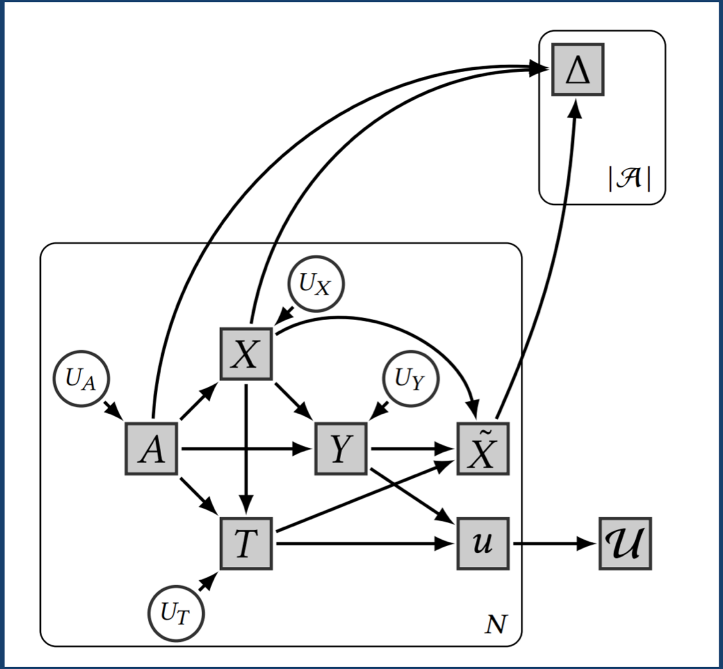
j-th Group avg score improvement Δ_j



^ Per-group score change for various bank policies

Liu et al 2018

Delayed Impact of Fair ML



Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j

Structural eqns:

Bank policy $T = f_T(U_T, A, X)$

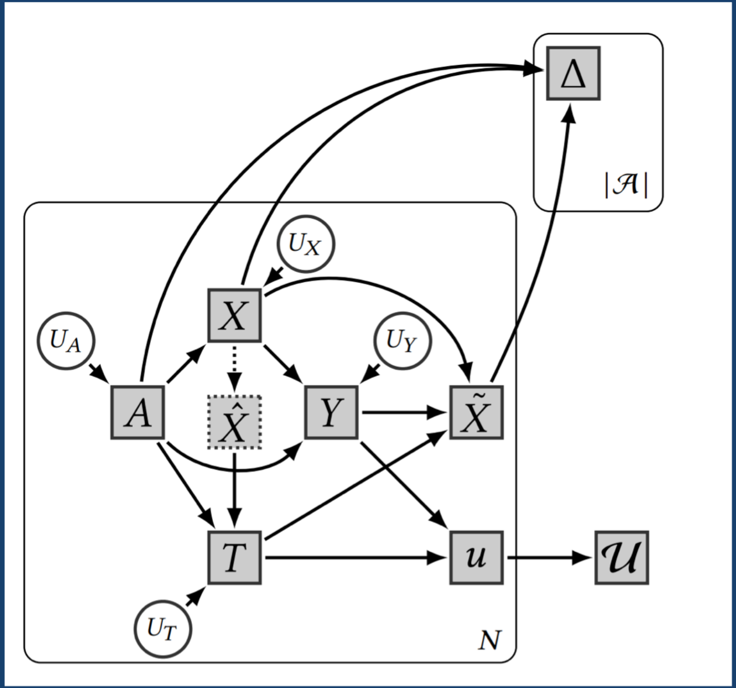
Potential outcome $Y = f_Y(U_Y, X, A)$

Next-step score $\tilde X = f_{\tilde X}(Y, T, X)$

j-th Group avg score improvement Δ_j

Liu et al 2018
Delayed Impact of Fair ML

Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j



Structural eqns:

Credit bureau policy \hat X = f_{\hat X}(X)

Bank policy ***T = f_T(U_T, A, X)***

Potential outcome ***Y = f_Y(U_Y, X, A)***

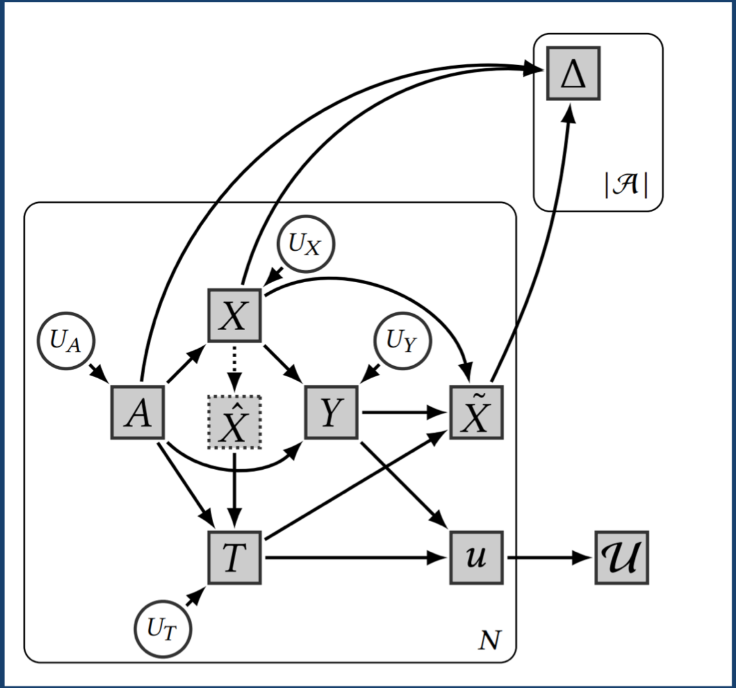
Next-step score ***\tilde X = f_{\tilde X}(Y, T, X)***

j-th Group avg score improvement ***\Delta_j***

Liu et al 2018

Delayed Impact of Fair ML

Institutional and group outcomes under double intervention ->



Structural eqns:

Credit bureau policy \ha

Bank policy ***T = f_T(U_T,***

Potential outcome ***Y = f_***

Next-step score ***\tilde X***

j-th Group avg score imp

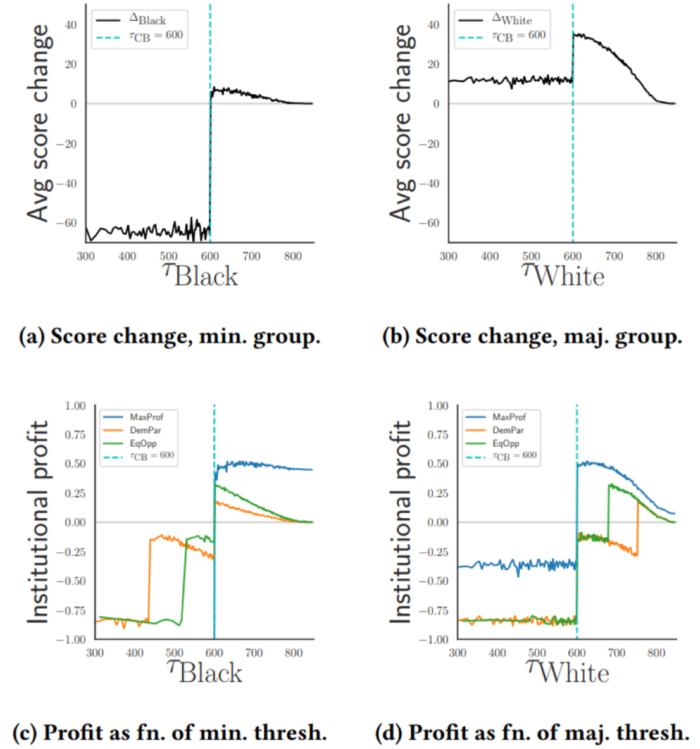
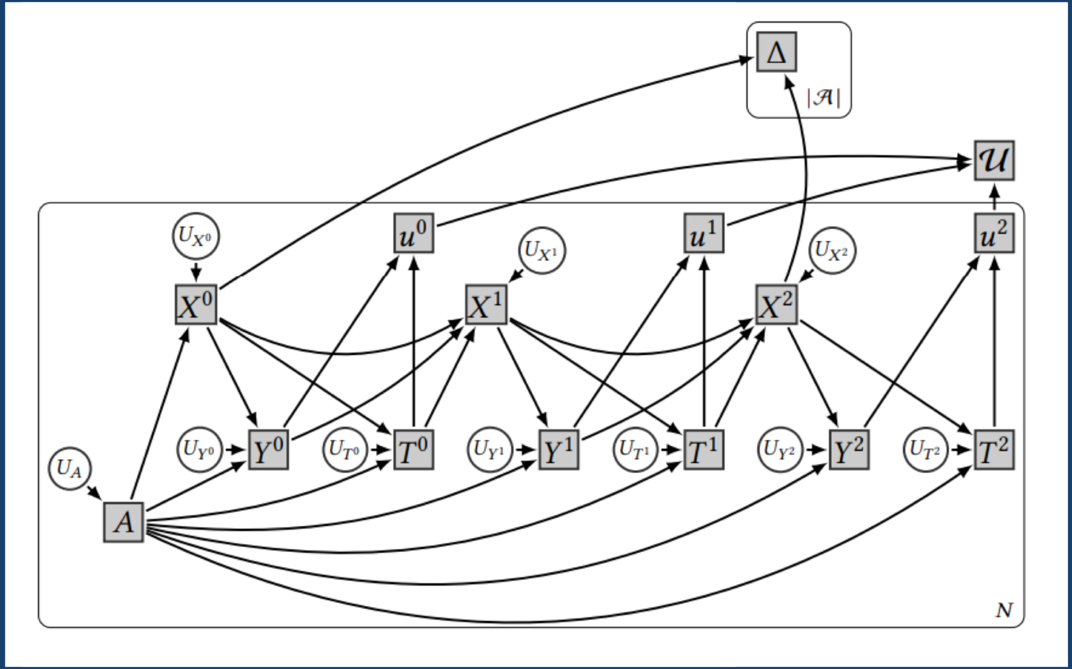


Figure 9: Policy evaluation under credit bureau intervention $\hat{f}_{\hat{X}}(X) = \min(X, \tau_{CB})$ with $\tau_{CB} = 600$. Group score change—formally $\mathbb{E}_{p^{do(f_{\hat{X}} \to \hat{f}_{\hat{X}}, f_T \to \hat{f}_T)}}[\Delta_j] \forall j \in \{\text{Black}, \text{White}\}$—and institutional profits—formally $\mathbb{E}_{p^{do(f_{\hat{X}} \to \hat{f}_{\hat{X}}, f_T \to \hat{f}_T)}}[\mathcal{U}]$—are shown as functions of the two group thresholds $\{\tau_j\}$. Bank profits depend on its fairness criteria.

Liu et al 2018
Delayed Impact of Fair ML

Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j



Multi-step structural eqns:

Bank policy ***T = f_T(U_T, A, X)***

Potential outcome ***Y = f_Y(U_Y, X, A)***

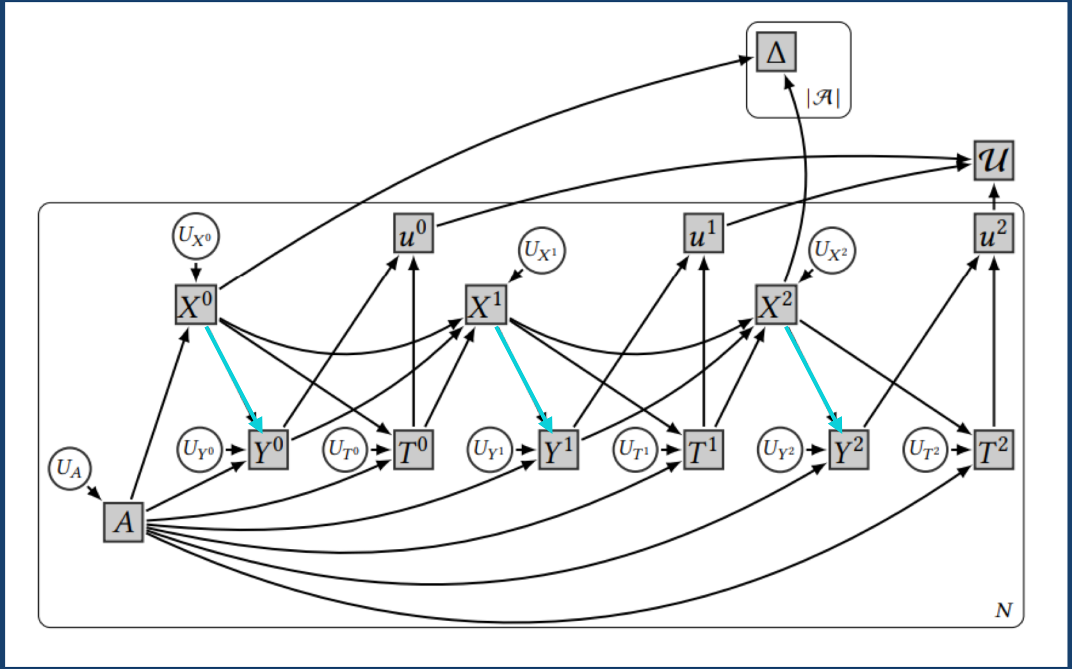
Next-step score ***\tilde X = f_{\tilde X}(Y, T, X)***

j-th Group avg score improvement ***\Delta_j***

Liu et al 2018

Delayed Impact of Fair ML

Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i
u_i	Utility of individual i (from the institution's perspective)
Δ_i	Expected improvement of score for individual i
$\tilde{X}_i$	Score for individual i after one time step
$\mathcal{U}$	Global utility (from institution's perspective)
Δ_j	Expected change in score for group j



Multi-step structural eqns:

Robustness intervention: f_Y -> f_{\hat Y}

Bank policy ***T = f_T(U_T, A, X)***

Potential outcome ***Y = f_Y(U_Y, X, A)***

Next-step score ***\tilde X = f_{\tilde X}(Y, T, X)***

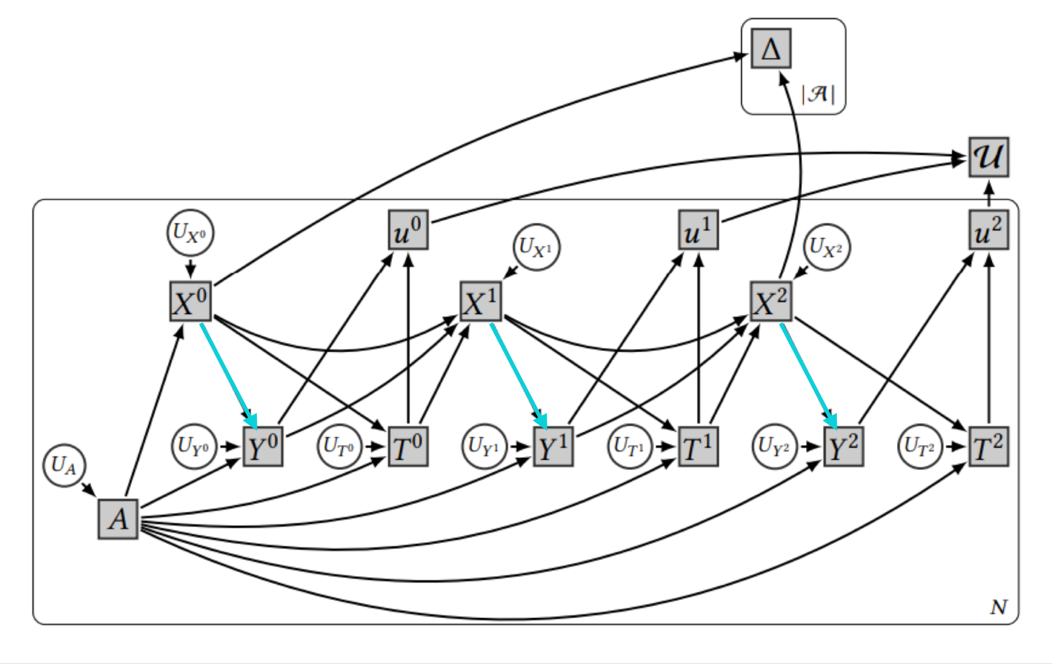
j-th Group avg score improvement ***\Delta_j***

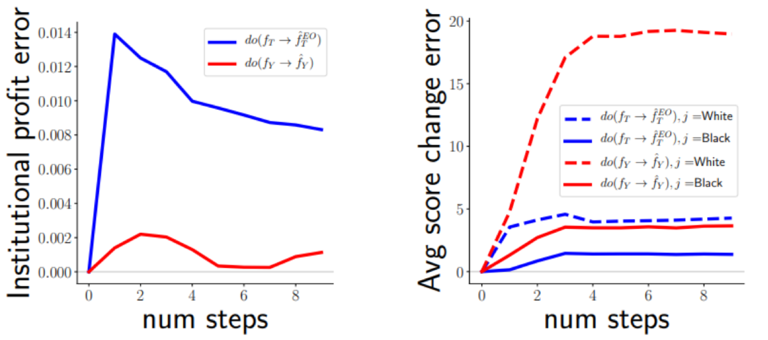
Liu et al 2018

Delayed Impact of Fair ML

Symbol	Meaning
N	Number of individuals
$\lvert\mathcal{A}\rvert$	Number of demographic groups
A_i	Sensitive attribute for individual i
U_{A_i}	Exogenous noise on sensitive attribute for individual i
X_i	Score for individual i
U_{X_i}	Exogenous noise on score for individual i
Y_i	Potential outcome (loan repayment/default) for individual i
U_{Y_i}	Exogenous noise on potential outcome for individual i
T_i	Treatment (institution gives/withholds loan) for individual i
U_{T_i}	Exogenous noise on treatment for individual i

Evaluating policy robustness via potential outcome intervention ->



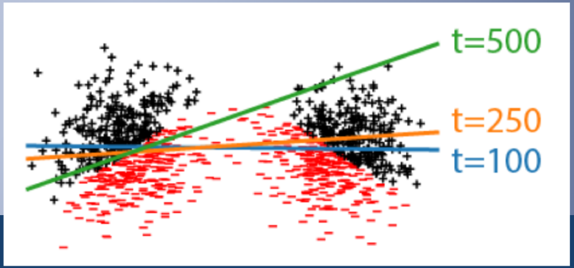


(a) Group improvement. **(b) Institutional profit.**

Figure 10: Evaluating multi-step policy robustness to distribution shift for various choice of intervention distribution q. Sensitivity of institutional utility—formally $|\mathbb{E}_q[\mathcal{U}] - \mathbb{E}[\mathcal{U}]|$—and sensitivity of group avg. score change—formally $|\mathbb{E}_q[\Delta_j] - \mathbb{E}[\Delta_j]|$—are shown as a function of steps. Expected profit is relatively robust to both interventions, whereas the expected per-group score changes are relatively more sensitive to these interventions.

Hashimoto et al 2018

Fairness w/o Demographics...



^ Population dynamics lead to classifier ignoring demographic minority

starting at $\lambda_k^{(0)} = b_k$ *is governed by:*

$$\lambda_k^{(t+1)} := \lambda_k^{(t)} \nu(\mathcal{R}_k(\theta^{(t)})) + b_k$$

$$\alpha_k^{(t+1)} := \frac{\lambda_k^{(t+1)}}{\sum_{k' \in [K]} \lambda_{k'}^{(t+1)}}$$

Demographic group mixture model

Group membership unobserved

Dynamics in group sizes

Evaluated learning via distributionally robust optimization

Structural eqns:

Latent group membership ***Z_i***

Mixture components ***(X_i,Y_i) = f_{(X_i,Y_i)}(Z_i == k, P_k)***

Learning algorithm \theta = f_\theta(U_\theta, {X_i, Y_i})

Predictions ***\hat Y_i = f_{\theta, \hat Y_i, X_i)***

Latent per-group risk ***R_i = f_R(Z_i == k, Y_i == \hat Y_i)***

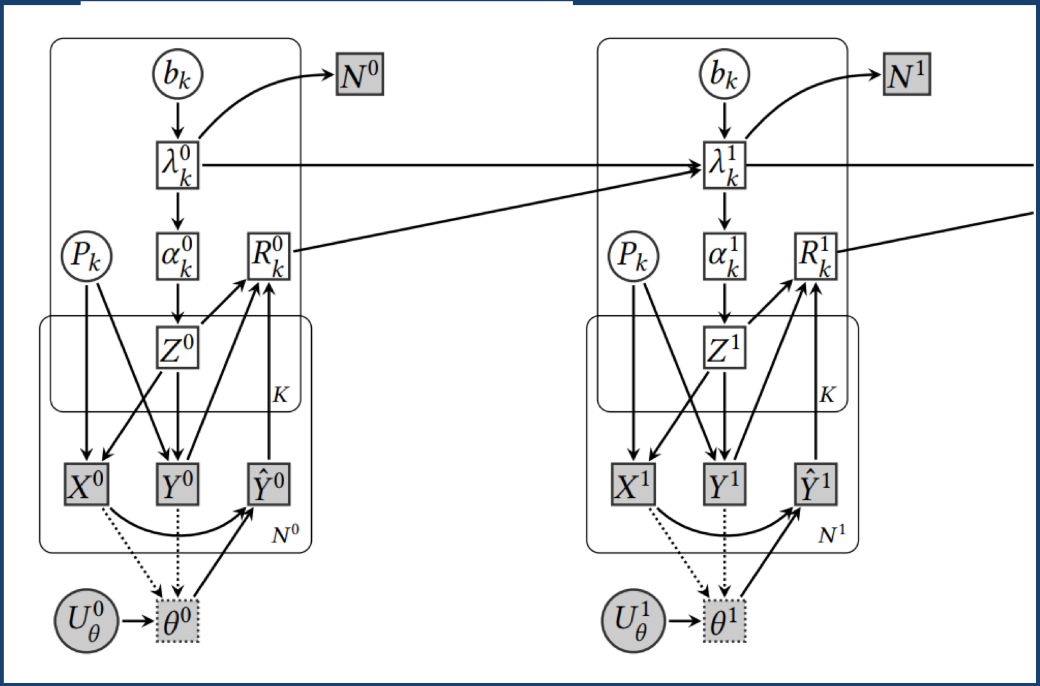
Latent group dynamics

\lamba_k^t+1 = f_\lambda(\lambda^t, R_k^t)

Hashimoto et al 2018

Fairness w/o Demographics...

k	indexes groups
P_k	distribution over (X, Y) for group k
b_k	expected group-k baseline population growth at each step
λ_k^t	expected population for group k at time t
α_k^t	mixing coeff for group k at time t
N^t	Total population at time t
Z_k^t	indicator of individual belonging to k-th group
X^t	input features for an individual at time t
Y^t	label for an individual at time t
U_θ^t	Exogenous noise in learning algo. (e.g., random seed)
θ^t	Estimated classifier parameters at time t
$\hat{Y}^t$	Predicted label for an individual at time t
R_k^t	Classification error for group k at time t (unobserved)



Latent group membership ***Z_i***

Mixture components ***(X_i,Y_i) = f_{(X_i,Y_i)}(Z_i == k, P_k)***

Learning algorithm \theta = f_\theta(U_\theta, {X_i, Y_i})

Predictions ***\hat Y_i = f_{\theta, \hat Y_i, X_i)***

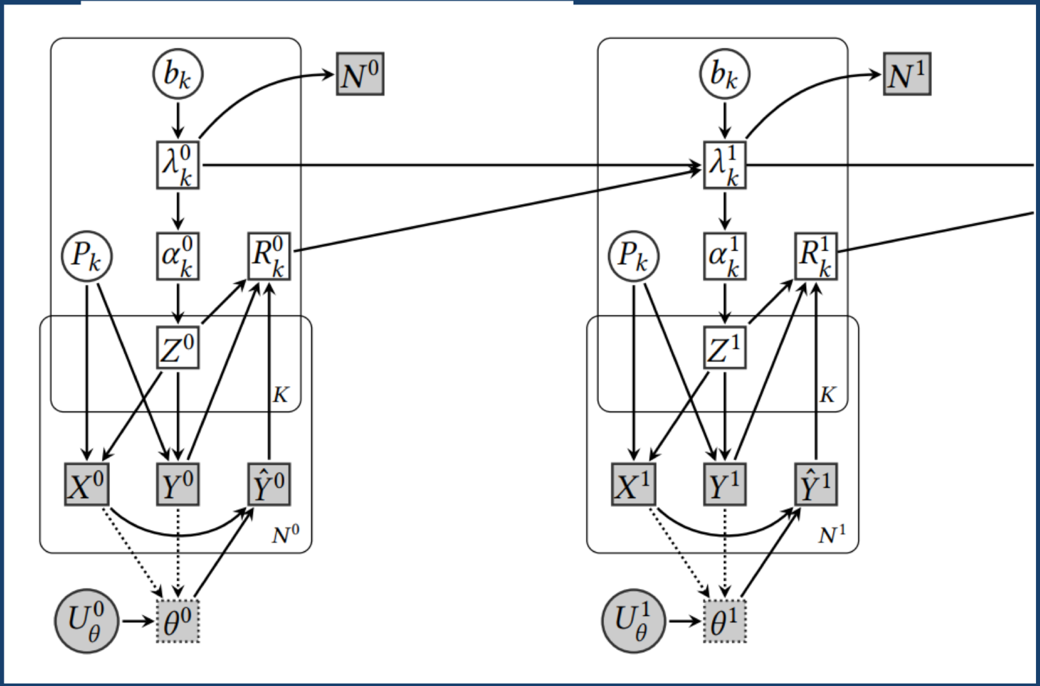
Latent per-group risk ***R_i = f_R(Z_i == k, Y_i == \hat Y_i)***

Latent group dynamics

\lamba_k^t+1 = f_\lambda(\lambda^t, R_k^t)

Hashimoto et al 2018

Fairness w/o Demographics...



k	indexes groups
P_k	distribution over (X, Y) for group k
b_k	expected group-k baseline population growth at each step
λ_k^t	expected population for group k at time t
α_k^t	mixing coeff for group k at time t
N^t	Total population at time t
Z_k^t	indicator of individual belonging to k-th group
X^t	input features for an individual at time t
Y^t	label for an individual at time t
U_θ^t	Exogenous noise in learning algo. (e.g., random seed)
θ^t	Estimated classifier parameters at time t
$\hat{Y}^t$	Predicted label for an individual at time t
R_k^t	Classification error for group k at time t (unobserved)

Latent group membership ***Z_i***

Mixtu)

Lea)

)

Lat)

Extensions of interest:

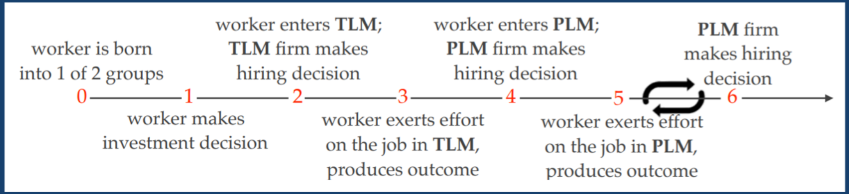
1. Intervene on group dynamics
2. Intervene on group distributions
3. Add dynamics to group distns
4. Off-policy evaluation:
   Can performance of a fair policy be estimated using trajectories recorded under a different policy?

\lamba_k^t+1 = f_\lambda(\lambda^t, R_k^t)

Hu and Chen 2018

A Short-term intervention...

TLM = temporary labor market, PLM = permanent labor market

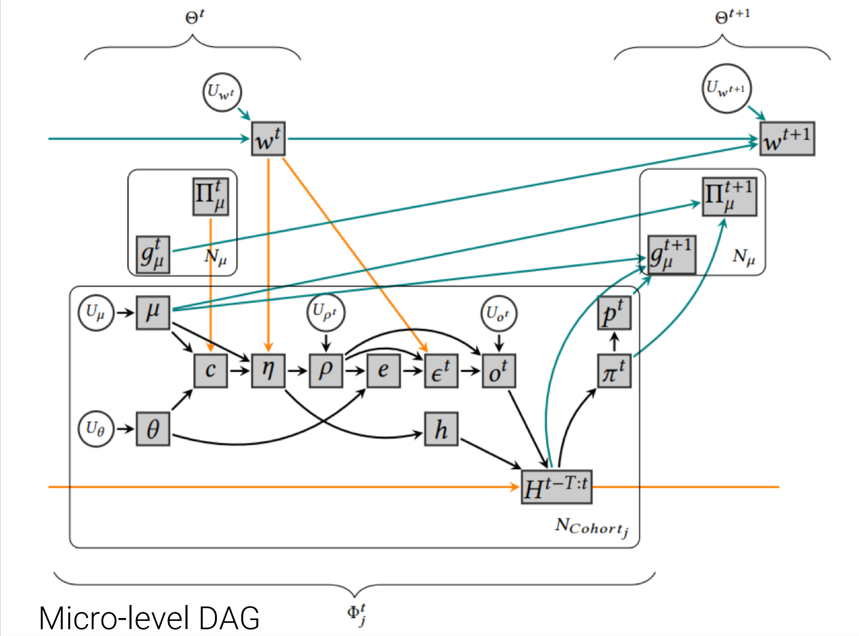


Models strategy of employees & employers

Hiring model with temporary and permanent workers

Evaluated effectiveness of intervention in short-term market

Symbol	Meaning
t	indexes time
i	indexes individuals
j	indexes cohorts
w^t	wages at time t
g^t_μ	proportion "good" group-μ workers in PLM
Π^t_μ	group μ reputation at time t
μ_i	group membership for worker i
θ_i	individual i ability
c_i	cost of investment for individual i
η_i	investment level for individual i
ρ_i	qualification level for individual i
e_i	individual-i cost of effort
ϵ^t_i	individual-i actual effort exerted at time t
o^t_i	individual-i outcome at time t
h_i	was individual hired to TLM following education?
$H_i^{t-\tau:t-1}$	individual-i τ-recent history (outcomes and TLM/PLM status)
π^t_i	individual i reputation at time t
p^t_i	was individual hired to PLM at step t?



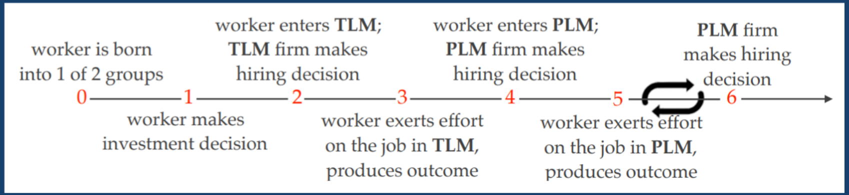
Micro-level DAG

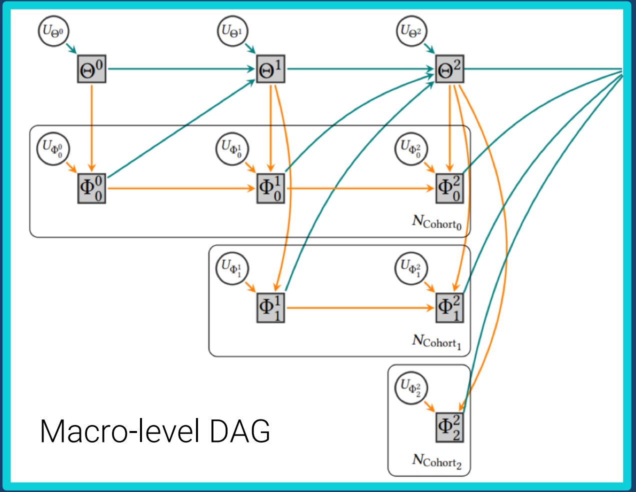
Hu and Chen 2018

A Short-term intervention...

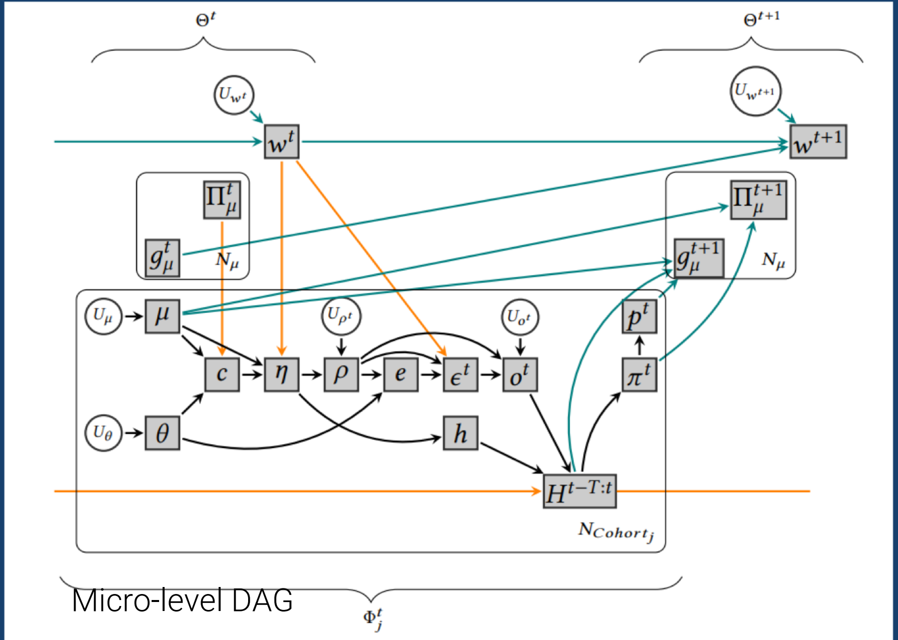
Symbol	Meaning
t	indexes time
i	indexes individuals
j	indexes cohorts
w^t	wages at time t
g_μ^t	proportion "good" group-μ workers in PLM
Π_μ^t	group μ reputation at time t
μ_i	group membership for worker i
θ_i	individual i ability
c_i	cost of investment for individual i
η_i	investment level for individual i
ρ_i	qualification level for individual i
e_i	individual-i cost of effort
ϵ_i^t	individual-i actual effort exerted at time t
o_i^t	individual-i outcome at time t
h_i	was individual hired to TLM following education?
$H_i^{t-\tau:t-1}$	individual-i τ-recent history (outcomes and TLM/PLM status)
π_i^t	individual i reputation at time t
p_i^t	was individual hired to PLM at step t?

TLM = temporary labor market, PLM = permanent labor market





Macro-level DAG

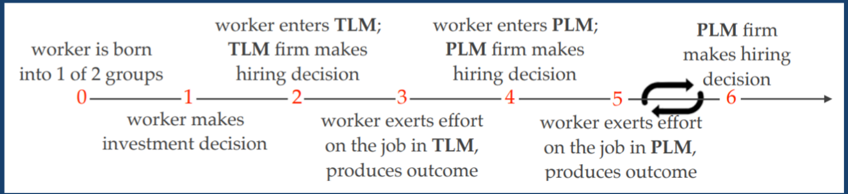


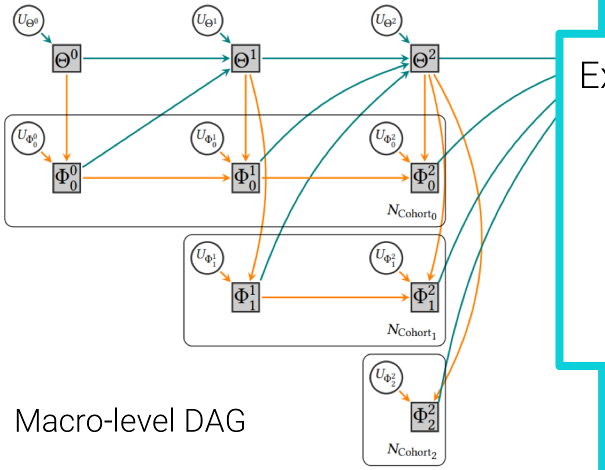
Micro-level DAG

Hu and Chen 2018

A Short-term intervention...

TLM = temporary labor market, PLM = permanent labor market



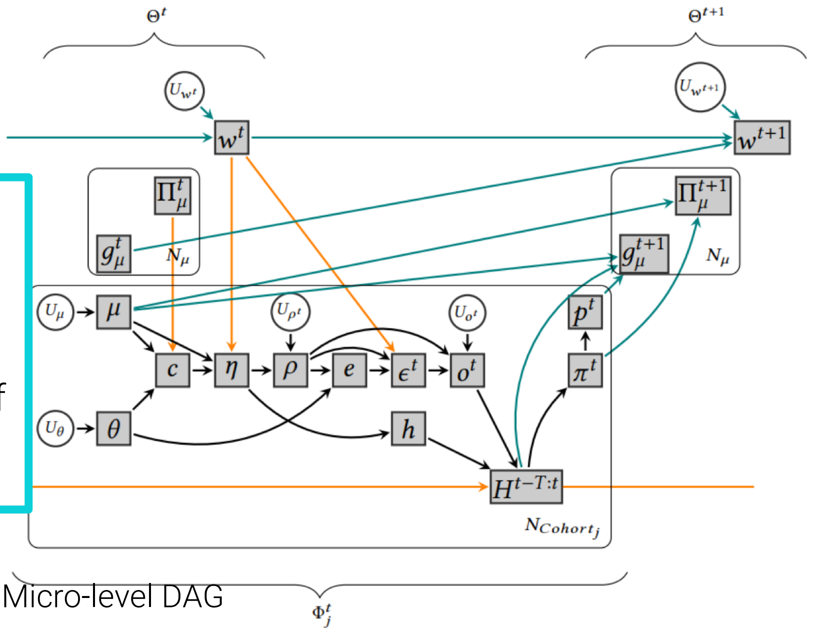


Macro-level DAG

Extensions of interest:

1. Intervene on cost of investment
2. Intervention on wages
3. Intervention on cost of effort

Symbol	Meaning
t	indexes time
i	indexes individuals
j	indexes cohorts
w^t	wages at time t
g_μ^t	proportion "good" group-μ workers in PLM
Π_μ^t	group μ reputation at time t
μ_i	group membership for worker i
θ_i	individual i ability
c_i	cost of investment for individual i
η_i	investment level for individual i
ρ_i	qualification level for individual i
e_i	individual-i cost of effort
ϵ_i^t	individual-i actual effort exerted at time t
o_i^t	individual-i outcome at time t
h_i	was individual hired to TLM following education?
$H_i^{t-\tau:t-1}$	individual-i τ-recent history (outcomes and TLM/PLM status)
π_i^t	individual i reputation at time t
p_i^t	was individual hired to PLM at step t?



Micro-level DAG

Summary

Causal DAGS are a unifying framework for recent work on long-term fairness

Causal DAGS enable :1. Visualization 2. Introspection 3. Evaluation

Some experimental procedures to consider:

- check robustness via interventions
  - models should exhibit robustness to some drift in test distribution
  - see also ***Invariant Risk Minimization*** [Arkjovsky et al 2019]

- off-policy evaluations
  - can we accurately estimate how new “fair” algorithms will perform in the real world?
  - see also counterfactual policy evaluation [Buesing et al 2018]

Future Work

- **Reinforcement learning and fairness:** Finding off-policy estimation methods better for low data, high-stakes regimes
- **Causal inference and dynamical systems:** Characterizing identifiability of long-term effects of policy interventions in terms of graphical criterion
- **Reinforcement learning and causal inference:** Developing methods for sensitivity analysis to estimate uncertainty of policy evaluations under confounding
- **Causal inference and visualization:** Visualizing complex, many-variable graphical models of policy problems
- **Fairness and decision science:** Integrating theoretical models of fairness in into scenario-based planning procedures