Accurate and Fair Machine Learning based on Causality

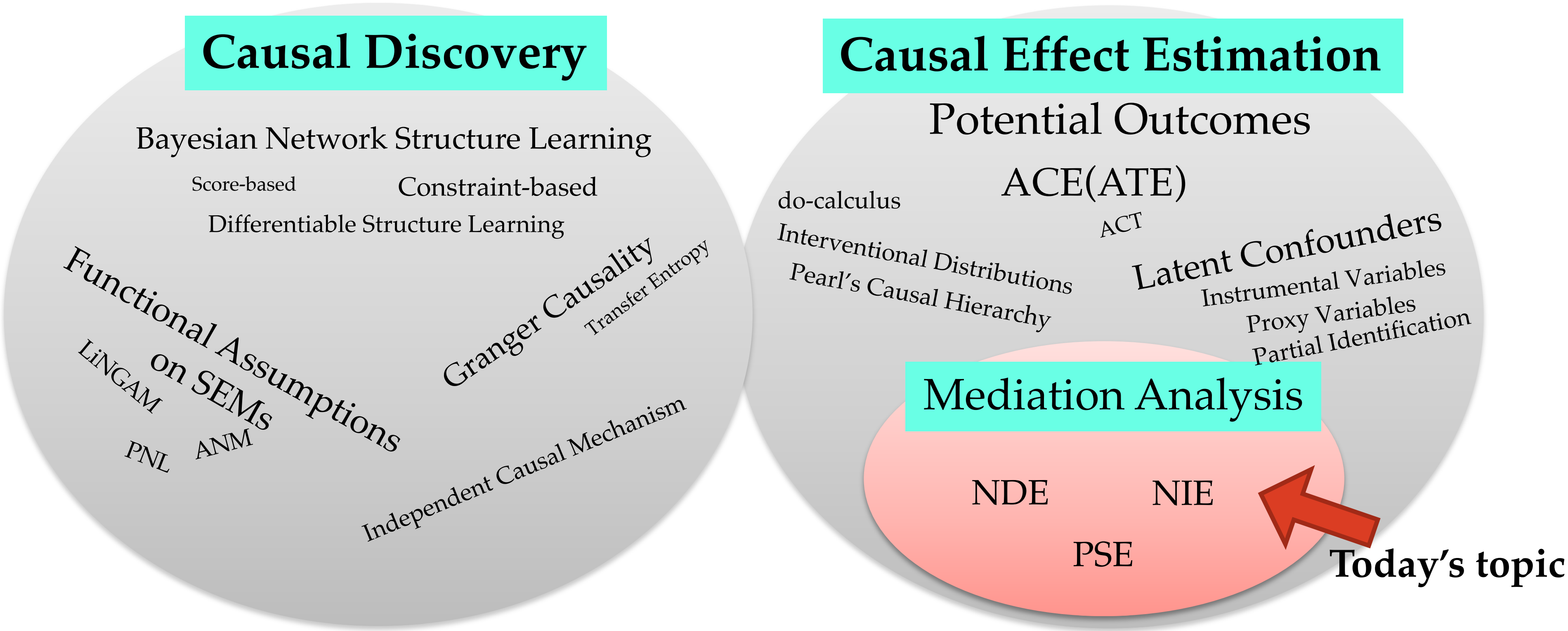
因果関係に基づく公平・高精度な機械学習の実現

NTT Communication Science Laboratories

Yoichi Chikahara

About Me

- **Name**: Yoichi Chikahara
- **Research Interest**: Causal Inference and ML.



X→Y, X←Y, or no causation?

How strong is this causality?
What happens if X's value is changed?

X→Y

Outline

1. Machine Learning and Fairness
    - Basic setup
    - Why do we need causality?
2. Introduction to Causal Effects
    - Potential outcomes, Average causal effect (ACE)
    - Mediation Analysis
3. Learning Fair Predictive Models based on Causality
    - Causality-based fairness criteria
    - Challenges: learn under weak assumptions

Outline

1. Machine Learning and Fairness
   - Basic setup
   - Why do we need causality?
2. Introduction to Causal Effects
   - Potential outcomes, Average causal effect (ACE)
   - Mediation Analysis
3. Learning Fair Predictive Models based on Causality
   - Causality-based fairness criteria
   - Challenges: learn under weak assumptions

Background to ML and fairness

- ML is increasingly used to make decisions for individuals

Application examples:
loan approval, job hiring, child abuse screening, and recidivism prediction

- Predicted decisions should be

Accurate **and** **Fair w.r.t. sensitive features** (e.g., gender, race, religion, disabilities, sexual orientation, etc.)

Problem setting example

Training a fair classifier

Training data

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Female	A	Economics	B	Accept
Male	B	Literature	B	Accept
Male	C	Science	C	Reject

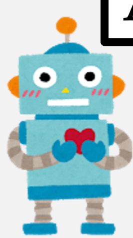
Solve constrained/penalized optimization problem

prediction loss

penalty on unfairness

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n),$$

Accurate & fair classifier

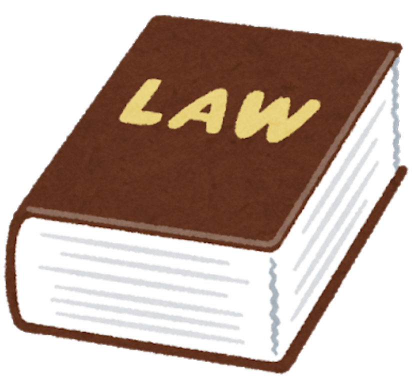


$$h_{\hat{\theta}}(A, Q, D, M)$$

Law defines the discrimination

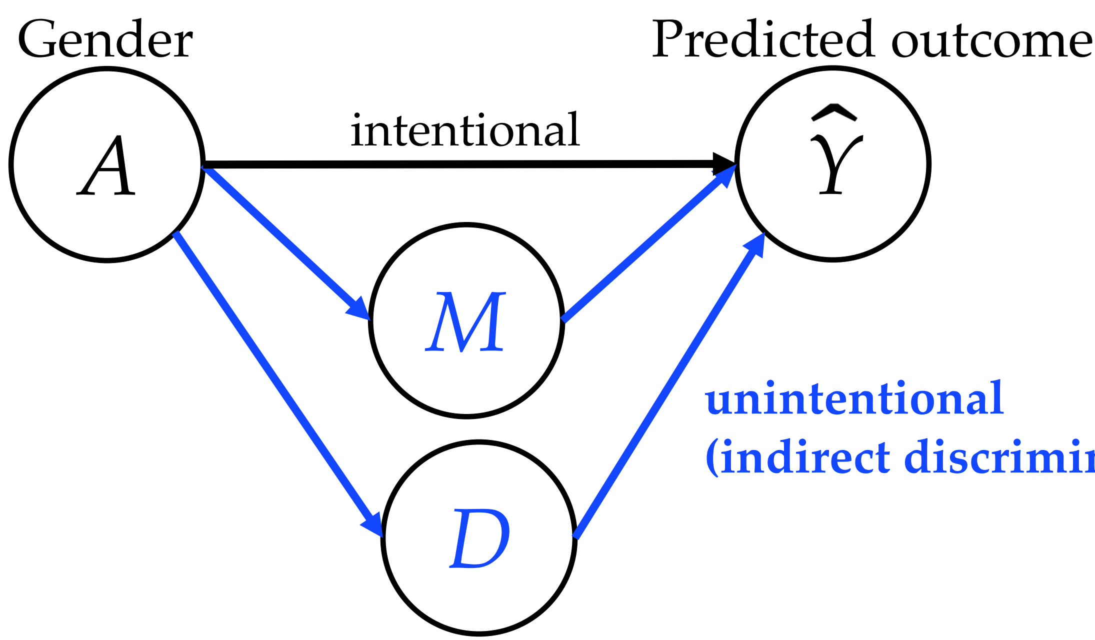
Example:

- **Disparate impact:**
    - Unintentional discrimination.
    - Even an apparently neutral policy should be prohibited if it adversely affects a privileged group (i.e., majority) more than unprivileged group (i.e., minority)
        - First defined by the U.S. Law called *Title VII of the 1964 Civil Rights Act*



How does unintentional discrimination occur?

- There are many *unintentional* factors that yield the correlation:
    - Use of features that are correlated with sensitive feature A



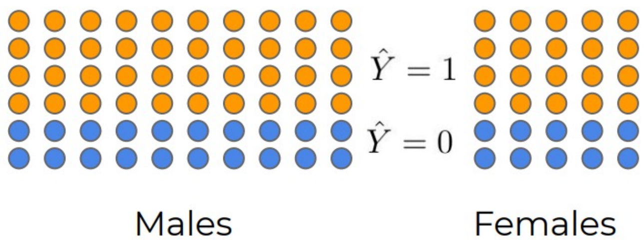
Fairness criteria for addressing disparate impact

Example:

Demographic parity (a.k.a., statistical parity):

- In binary classification, **the percentage of individuals assigned to class 1 should be identical:**

$$\mathrm{P}\big(\hat{Y} = 1 \big| A = 0\big) = \mathrm{P}(\hat{Y} = 1 | A = 1)$$



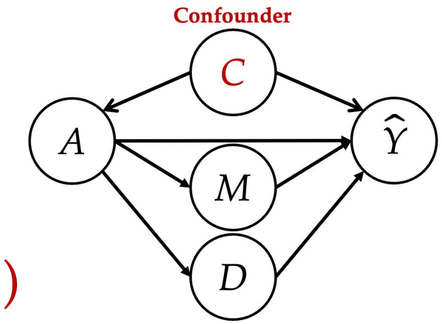
- In general, demographic parity requires independence between prediction $\hat{Y}$ and sensitive feature A

$$\hat{Y} \perp\!\!\!\perp A$$

  - For instance, HSIC [Gretton+; 2005] can be used to measure the independence

Weakness of correlation-based fairness criteria

1. No correlation does not imply no causation



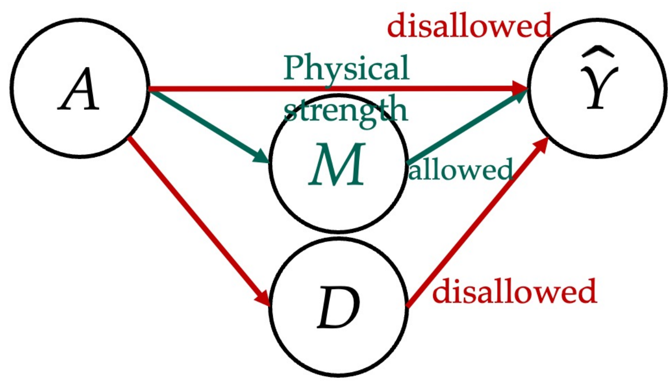
- Correlation between A and $\hat{Y}$ is determined by
  1. Causation from A to $\hat{Y}$ ($A \rightarrow \ldots \rightarrow \hat{Y}$)
  2. Confounding bias due to confounder C ($A \leftarrow C \rightarrow \hat{Y}$)

- This indicates that **even when there is no correlation, sensitive feature A may have causal effects on outcome $\hat{Y}$ (i.e., no correlation does not imply no causation)**
  - This is a serious issue because discrimination claims in the Laws are judged based on causality ☹

Weakness of correlation-based fairness criteria

2. Cannot address scenarios with *allowed indirect discrimination*

- In real-world scenarios, **several types of indirect discrimination might be allowed**.
  - Example: To make hiring decisions for physically demanding jobs, indirect effects through physical strength M may be legally allowed.



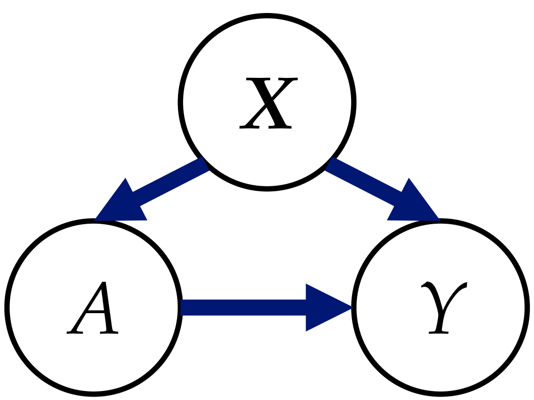
- In this case, **imposing no correlation is an unnecessarily restrictive fairness constraint.**
  - This is problematic because our goal is to achieve a tradeoff between fairness and accuracy ☹

Outline

1. Machine Learning and Fairness
   - Basic setup
   - Why do we need causality?
2. Introduction to Causal Effects
   - Potential outcomes, Average causal effect (ACE)
   - Mediation Analysis
3. Learning Fair Predictive Models based on Causality
   - Causality-based fairness criteria
   - Challenges: learn under weak assumptions

How can we measure *unfair* causal effects from the observed data?

Causal graph



Observed data

sensitive feature → A

observed outcome → Y

i	A	X_1	…	X_d	Y
1	1		…		Accept
2	0		…		Reject
3	0		…		Accept
4	1		…		Reject
5	1		…		Reject

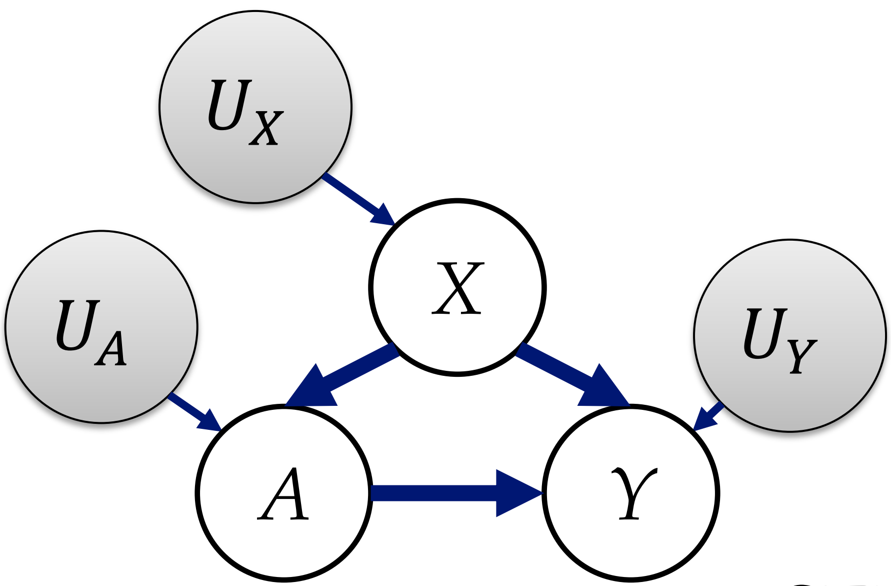
Basic notions

- **Potential outcome** Y(a)
  - Outcome Y that is observed when sensitive feature is A=a
  - $Y = aY(1) + (1-a)Y(0) \quad$ for $a \in \{0, 1\}$
- **Causal effect (a.k.a., treatment effect) for an individual:**
  - Difference between potential outcomes: Y(1) – Y(0)
  - **Can never be observed**

i	A	X_1	…	X_d	Y	Y(1)	Y(0)	Y(1) - Y(0)
1	1		…		Accept	Accept	?	?
2	0		…		Reject	?	Reject	?
3	0		…		Accept	?	Accept	?
4	1		…		Reject	Reject	?	?
5	1		…		Reject	Reject	?	?

How are potential outcomes defined?

- **A structural equation model (SEM)** [Pearl; 2000] contains
  - Observed variables (a.k.a., endogenous variables): A, X, Y
  - Unobserved noise variables (a.k.a., exogenous variables): U_A, U_X, U_Y
  - Deterministic functions: f_A, f_X, f_Y



Structural equations:

$$X = f_X(U_X)$$

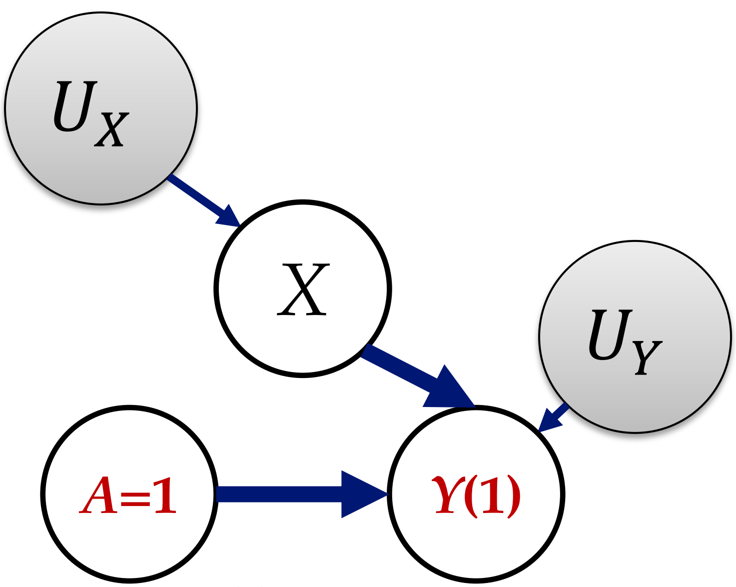
$$A = f_A(X, U_A)$$

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

SEM M

How are potential outcomes defined with an SEM?

- Definition: **Potential outcome is outcome Y in a different SEM whose structural equation of A is replaced.**
  - Such a replacement of structural equations is called *intervention do(A=a)*



$$X = f_X(U_X)$$

$$A = 1$$

$$Y(1) = f_Y(1, X, U_Y)$$

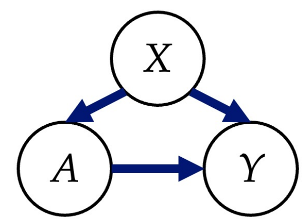
Interventional SEM $M_{do(A=1)}$

Average of causal effects can be estimated

- Average causal effect (ACE) across individuals **can be estimated**

i	A	X_1	…	X_d	Y	Y(1)	Y(0)	Y(1) - Y(0)
1	1		…		Accept	Accept	?	?
2	0		…		Reject	?	Reject	?
3	0		…		Accept	?	Accept	?
4	1		…		Reject	Reject	?	?
5	1		…		Reject	Reject	?	?

Average can be estimated



- **Note:** $\mathrm{E}[Y(a)] \neq \mathrm{E}[Y|A = a]$
  - E.g., $\mathrm{E}[Y(1)] \neq \mathrm{E}[Y|A = 1]$
  - Why? Because group A=0 and group A=1 often **have different attributes X**. Taking average over different groups does not make sense.

Example:

› **Age**

old old old old old young old

young young old young young old old young

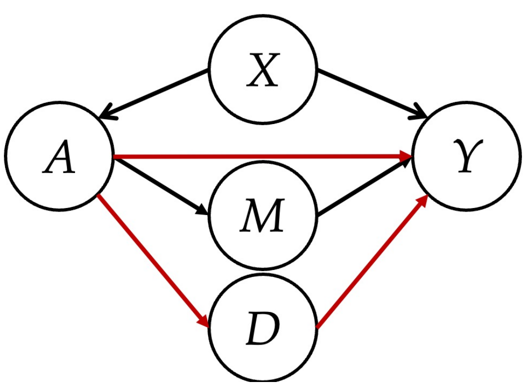
$A = 1$ (Has prior conviction) $A = 0$ (No prior conviction)

ACE Estimation

- <u>*Ignorability* Assumption</u>: Features $\boldsymbol{X}$ contains all *confounders*
  - Formally, $A \perp\!\!\!\perp Y(a) \mid \boldsymbol{X}$ holds for any $a \in \{0, 1\}$
- Under this assumption, ACE can be estimated by
  - *g-formula*:
    - $\mathrm{E}[Y(1) - Y(0)] = \sum_{\boldsymbol{X}}(\mathrm{E}[Y|A = 1, \boldsymbol{X}] - \mathrm{E}[Y|A = 0, \boldsymbol{X}])\mathrm{P}(\boldsymbol{X})$
  - *Inverse probability weighting* (IPW)
    - Importance sampling technique for computing an expected value w.r.t. $\mathrm{P}(\boldsymbol{X})$ using samples from $\mathrm{P}(\boldsymbol{X}|A = a)$
    - $\mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[\frac{a}{\mathrm{P}(A=1|X)}Y]$-$\mathrm{E}[\frac{1-a}{1-\mathrm{P}(A=1|X)}Y]$

How can we measure causal effects along *unfair pathways*?

Complicated causal graph

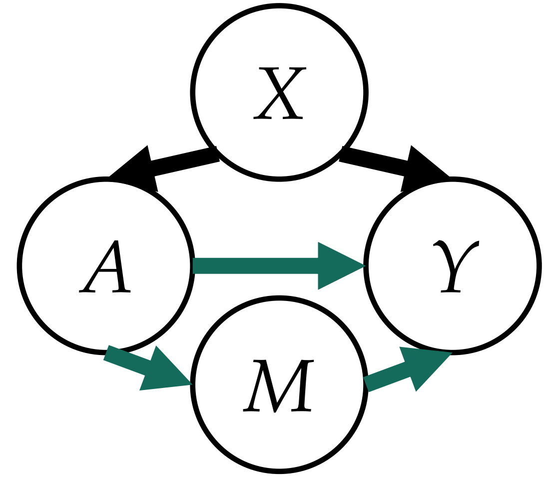


Observed data

sensitive feature → A; observed outcome → Y

i	A	D	M	X	Y
1	1				Accept
2	0				Reject
3	0				Accept
4	1				Reject
5	1				Reject

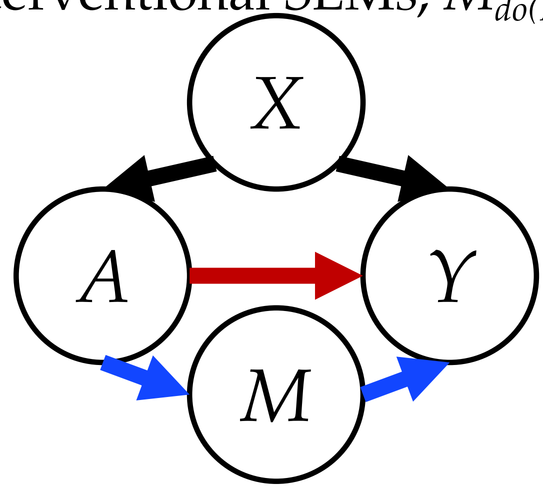
How can we measure causal effects along pathways?



- Consider causal graph with mediator M
  - Mediator M is also affected by A
  - Outcome Y is influenced by A and M
- Potential mediators M(a)
  - Mediator M that is observed when sensitive feature is A=a
  - $M = aM(1) + (1 - a)M(0)$ for $a \in \{0, 1\}$
- Using potential mediators, causal effect for an individual is formulated as
  - Y(1, M(1)) – Y(0, M(0))
- This causal effect corresponds to a total causal effect along all pathways from $A \rightarrow Y$

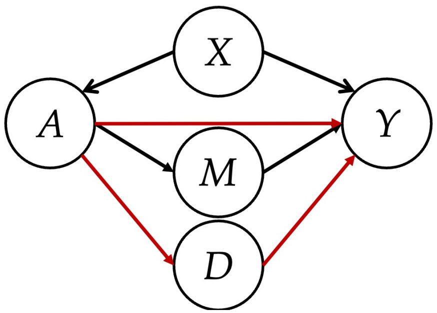
Direct effects and Indirect effects

- Using potential mediators, we can also measure causal effects along direct and indirect pathways, i.e., a natural direct effect (NDE) and a natural indirect effect (NIE):
  - $NDE = Y(\underline{1,}\, M(0)) - Y(\underline{0,}\, M(0))$
  - $NIE = Y(0, M(\underline{1})) - Y(0, M(\underline{0}))$
    - *Note:* Nested potential outcomes $Y(0, M(1))$ and $Y(1, M(0))$ are defined with two interventional SEMs, $M_{do(A=0)}$ and $M_{do(A=1)}$



Path-specific causal effects (PSE) [Avin+; IJCAI2005]

- Consider more complicated causal graph with multiple mediators



- Causal effects along pathways π ={A→Y, A→D→Y} are measured by path-specific causal effects (PSE) [Avin+; IJCAI2005] as
  - $PSE(\pi) = Y(1 \parallel \pi) - Y(0)$
    - › $Y(1 \parallel \pi) \equiv Y(\underline{1,} D(\underline{1}), M(0))$
    - › $Y(0) \equiv Y(0, D(0), M(0))$

Changed to A=1 if the variable is a node in pathway set π

- Mean potential outcome $\mathrm{E}[Y_{A \Leftarrow 1 \parallel \pi}]$ can be similarly computed by
  - *Edge-g-formula* [Shpitser+; AS2015]
  - *Inverse probability weighting*

Outline

1. Machine Learning and Fairness
   - Basic setup
   - Why do we need causality?
2. Introduction to Causal Effects
   - Potential outcomes, Average causal effect (ACE)
   - Mediation Analysis
3. Learning Fair Predictive Models based on Causality
   - Causality-based fairness criteria
   - Challenges: learn under weak assumptions

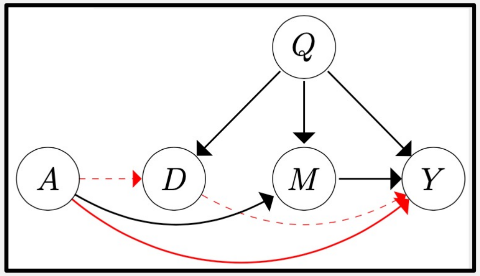
Problem setting

Training a fair classifier with causal graph

Training data

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Female	A	Economics	B	Accept
Male	B	Literature	B	Accept
Male	C	Science	C	Reject

Causal graph



- Given by prior domain knowledge
- Inferred by causal discovery algorithm

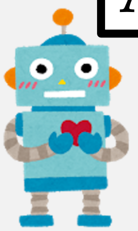
Unfair pathways
$\pi = \{A \to Y, A \to D \to Y\}$

Solve constrained/penalized optimization problem

prediction loss **penalty on unfairness**

$$\min_{\theta} \quad \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n),$$

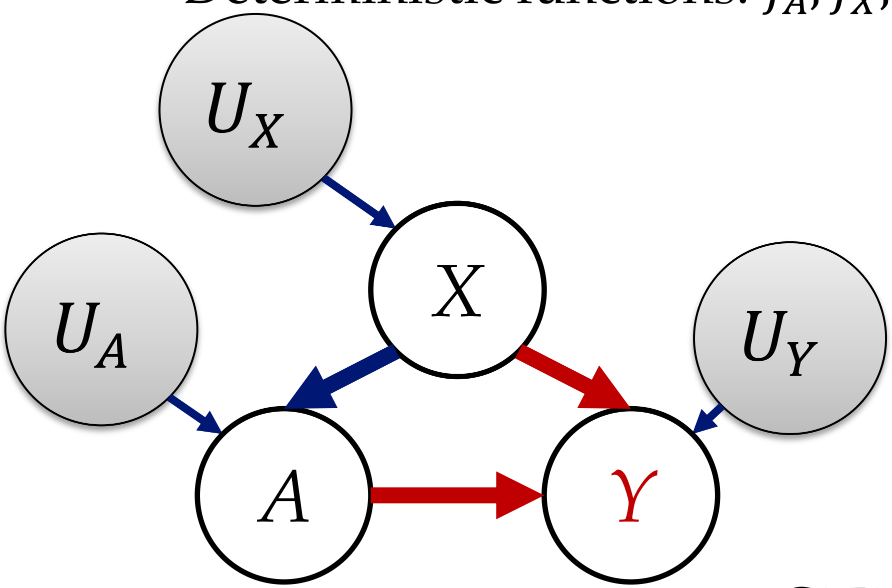
Accurate & fair classifier



$h_{\hat{\theta}}(A, Q, D, M)$

Potential outcomes for prediction

- To formulate potential outcomes for prediction Y, we consider a little bit different SEM:
  - Observed variables (a.k.a., endogenous variables): A, X, Y
  - Unobserved noise variables (a.k.a., exogenous variables): U_A, U_X, U_Y
  - Deterministic functions: f_A, f_X, h_θ



SEM M^p

Structural equations:

$$X = f_X(U_X)$$

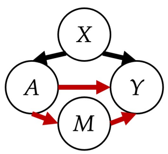
$$A = f_A(X, U_A)$$

$$Y = h_\theta\,(A, X, U_Y)$$

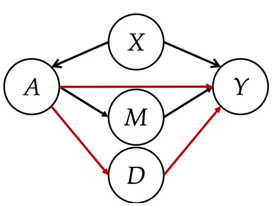
Prediction Y is determined by classifier h_θ

Causality-based fairness criteria

	Group-level	Individual-level
Total effects	Fair on ACE (**FACE**) [Khademi+; WWW2019] ACE: $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = 0$	**Counterfactual fairness** [Kusner+; NeurIPS2017 Best Paper] $\mathrm{E}[Y(1) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] - \mathrm{E}[Y(0) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = 0$ for all a and $\boldsymbol{x}$
Path-specific effects	**Path-specific population-level fairness** [Nabi+; AAAI2018] PSE: $\mathrm{E}[Y(1 \parallel \pi)] - \mathrm{E}[Y(0)] = 0$	**Path-specific counterfactual fairness (PC-fairness)** [Wu+; NeurIPS2019] $\mathrm{E}[Y(1 \parallel \pi) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] - \mathrm{E}[Y(0) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = 0$ for all a and $\boldsymbol{x}$



All pathways from A to Y are unfair

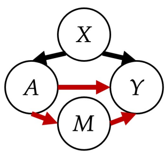


We can choose unfair pathways

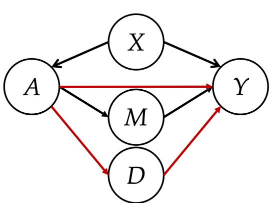
<u>Note</u>: For simplicity, Y is regarded as binary

Causality-based fairness criteria

	Group-level	Individual-level
Total effects	Fair on ACE (**FACE**) [Khademi+; WWW2019] ACE: $\mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = 0$	**Counterfactual fairness** [Kusner+; NeurIPS2017 Best Paper] $\mathrm{E}[Y(1) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] - \mathrm{E}[Y(0) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = 0$ for all a and $\boldsymbol{x}$
Path-specific effects	**Path-specific population-level fairness** [Nabi+; AAAI2018] PSE: $\mathrm{E}[Y(1 \parallel \pi)] - \mathrm{E}[Y(0)] = 0$	**Path-specific counterfactual fairness (PC-fairness)** [Wu+; NeurIPS2019] $\mathrm{E}[Y(1 \parallel \pi) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] - \mathrm{E}[Y(0) \mid A = a, \boldsymbol{X} = \boldsymbol{x}] = 0$ for all a and $\boldsymbol{x}$



All pathways from A to Y are unfair



We can choose unfair pathways

Note: For simplicity, Y is regarded as binary

Group-level fairness:
Remove the mean PSE [Nabi+; AAAI2018]

- Constrain average PSE across **all** individuals:

Prediction by h_θ

i	A	D	M	Q	Y	$Y(1 \| \pi)$	$Y(0)$	$Y(1 \| \pi)$-$Y(0)$
1	1	0	A	C	1	?	?	?
2	0	1	B	B	0	?	0	?
3	0	1	B	B	1	?	1	?
4	1	2	C	A	0	?	?	?
5	1	3	C	B	0	?	?	?

Average PSE on prediction

$$\mathrm{E}[Y(1 \parallel \pi)] - \mathrm{E}[Y(0)] = 0$$

Group-level fairness:
Remove the mean PSE [Nabi+; AAAI2018]

- However, removing the mean PSE does **not imply that predictions are fair for each individual**

i	A	D	M	Q	Y	$Y(1\|\pi)$	$Y(0)$	$Y(1\|\pi)$-$Y(0)$
1	1	0	A	C	1	?	?	**1**
2	0	1	B	B	0	?	0	**-1**
3	0	1	B	B	1	?	1	**1**
4	1	2	C	A	0	?	?	**-1**
5	1	3	C	B	0	?	?	**0**

Average PSE is zero, but **some individuals suffer from discrimination**

Individual-level fairness: Remove mean PSE for each subgroup [Chiappa+; AAAI2019]

- Separate individuals into subgroups with identical attributes of sensitive feature A and non-sensitive features $\boldsymbol{X}$
- Remove the mean PSE for each subgroup

i	A	D	M	Q	Y	$Y(1\|\pi)$	$Y(0)$	$Y(1\|\pi)$-$Y(0)$
1	1	0	A	C	1	?	?	**0**
2	0	1	B	B	0	?	0	**-1**
3	0	1	B	B	1	?	1	**1**
4	1	2	C	A	0	?	?	**0**
5	1	3	C	B	0	?	?	**0**

Attributes A and $\boldsymbol{X}$ are identical

Average PSE is zero **for each subgroup of individuals**

- Formally, this fairness criterion (PC-fairness [Wu+; NeurIPS2019]) is defined as

$$\mathrm{E}[Y(1 \parallel \pi)|A = a, \boldsymbol{X} = \boldsymbol{x}] - \mathrm{E}[Y(0)|A = a, \boldsymbol{X} = \boldsymbol{x}] = 0$$

for all a and $\boldsymbol{x}$

Weakness of existing methods for achieving PC-fairness

- <u>Issue</u>: **Conditional expectation of PSEs are difficult to estimate** (due to conditioning on mediators ☹)
  - Existing methods aim to approximate the true SEM; however, this approximation **requires a restrictive functional assumption on the SEM** ☹

Structural equations:

$$X = f_X(U_X)$$

$$A = f_A(X, U_A)$$

$$Y = h_\theta\,(A, X, U_Y)$$

These structural equations are assumed to be expressed as *additive noise model (ANM)*

$$V = f\big(\boldsymbol{pa}(\boldsymbol{V})\big) + U_V$$

However, it is unclear whether such an assumption holds ☹

Learning individually fair classifier with path-specific causal-effect constraint [Chikahara+; AISTATS2021]

<u>Our proposal</u>: Impose a constraint on the following probability:

- *Probability of Individual Unfairness* (PIU) [Chikahara+; AISTATS2021]

$$\text{PIU:} \quad \mathrm{P}(Y(0) \neq Y(1 \parallel \pi))$$

- **This joint probability can be never inferred** (because we can never jointly obtain potential outcomes $Y(0)$ and $Y(1 \parallel \pi)$)
- However, **upper bound on PIU can be estimated without making restrictive functional assumptions on the SEM** ☺

$$\mathrm{P}(Y(0) \neq Y(1 \parallel \pi)) \leq \underline{2\mathrm{P}^I(Y(0) \neq Y(1 \parallel \pi))}$$

$$= \underline{2(\mathrm{P}(Y(0)=1)\big(1-\mathrm{P}(Y(1 \parallel \pi)=1)\big) + (1-\mathrm{P}(Y(0)=1)\mathrm{P}(Y(1 \parallel \pi)=1))}$$

Y is binary

P^I: independent joint distribution of potential outcomes

Learning individually fair classifier with path-specific causal-effect constraint [Chikahara+; AISTATS2021]

- **Zero PIU is sufficient to guarantee PC-fairness ☺**
- So we formulate our penalty function G_θ using the estimator of upper bound on PIU:

$$\min_\theta \quad \frac{1}{n}\sum_{i=1}^{n} L_\theta(\boldsymbol{x}_i, y_i) + \lambda G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n),$$

$$G_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \hat{p}_\theta^{A \Leftarrow 1 \| \pi}(1 - \hat{p}_\theta^{A \Leftarrow 0}) + (1 - \hat{p}_\theta^{A \Leftarrow 1 \| \pi})\hat{p}_\theta^{A \Leftarrow 0}$$

where $\hat{p}_\theta^{A \Leftarrow 0}$ and $\hat{p}_\theta^{A \Leftarrow 1 \| \pi}$ are IPW–based estimators of $\mathrm{P}(Y(0) = 1)$ and $\mathrm{P}(Y(1 \parallel \pi) = 1)$; for instance,

$$\hat{p}_\theta^{A \Leftarrow 0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(a_i = 0)\hat{w}_i c_\theta(a_i, q_i, d_i, m_i) \qquad \hat{p}_\theta^{A \Leftarrow 1 \| \pi} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(a_i = 1)\hat{w}'_i c_\theta(a_i, q_i, d_i, m_i)$$

Learning individually fair classifier with path-specific causal-effect constraint [Chikahara+; AISTATS2021]

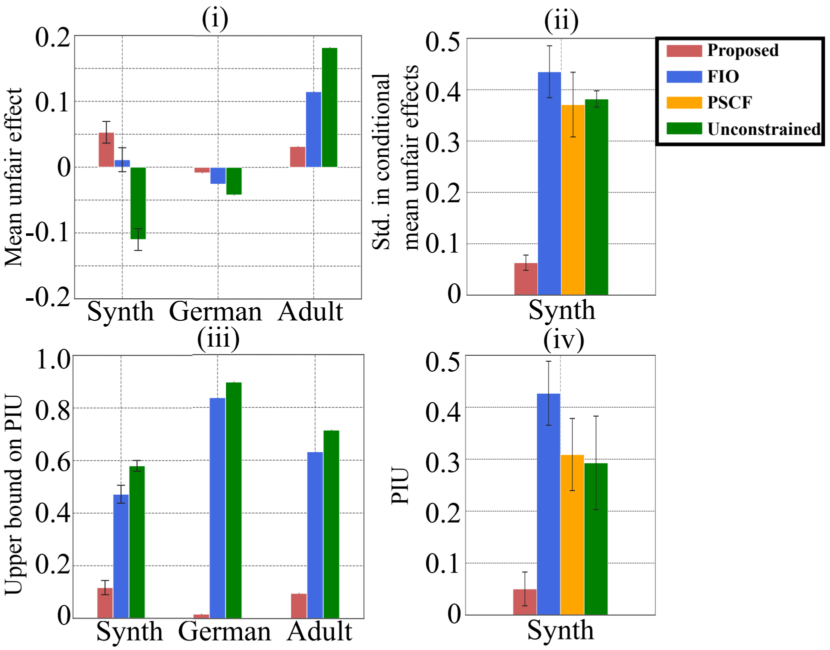
- **Proposed method experimentally strikes a good balance between accuracy and fairness ☺**

Table 2: Test accuracy (%) on each dataset

Method	Synth	German	Adult
Proposed	80.0 ± 0.9	75.0	75.2
FIO	84.8 ± 0.6	78.0	81.2
PSCF	74.8 ± 1.6	76.0	73.4
Unconstrained	88.2 ± 0.9	81.0	83.2
Remove	76.9 ± 1.3	73.0	74.7

Proposed (Red one) can eliminate unfair PSE for each individual ☺

Figure 2: Four statistics of unfairness on test data



There are many open problems and challenges

Take-home messages: Causality-based fairness is powerful, but causal inference requires assumptions. This makes it challenging to develop practical causality-based framework.

- Uncertain causal graph structure:
  - Multi-World Fairness (MWF) [Russell+; NeurIPS2017] uses multiple candidates of causal graphs
- Unidentifiable setting:
  - When there are unobserved confounders
    - Proxy variables, partial identification, etc. are helpful

Dealing with such settings remains an open problem

Conclusion

- Law defines discrimination. How do we measure it?
- Causality-based fairness can detect confounding bias
- Mediation analysis is helpful to strike a good balance between prediction accuracy and fairness
- There are many challenging open problems.

Thank you!