Lecture 19: Causal Fairness

Instructor: Jing Ma

Fall 2024, CDS@CWRU

CSDS 452 Causality and Machine Learning

Outline

- Fairness in machine learning
  - Group fairness, individual fairness
  - Causal fairness
- Achieving counterfactual fairness
  - 3-step method
  - Generative models

Biases in Machine Learning

Real-world inequality and discrimination lead to biases in machine learning

October 11, 2018

Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

By Roberto Iriondo

Algorithmic Fairness

Then how to define fairness?

Fairness can be defined in different ways [1]: different real-world applications show biases from various perspectives [2].

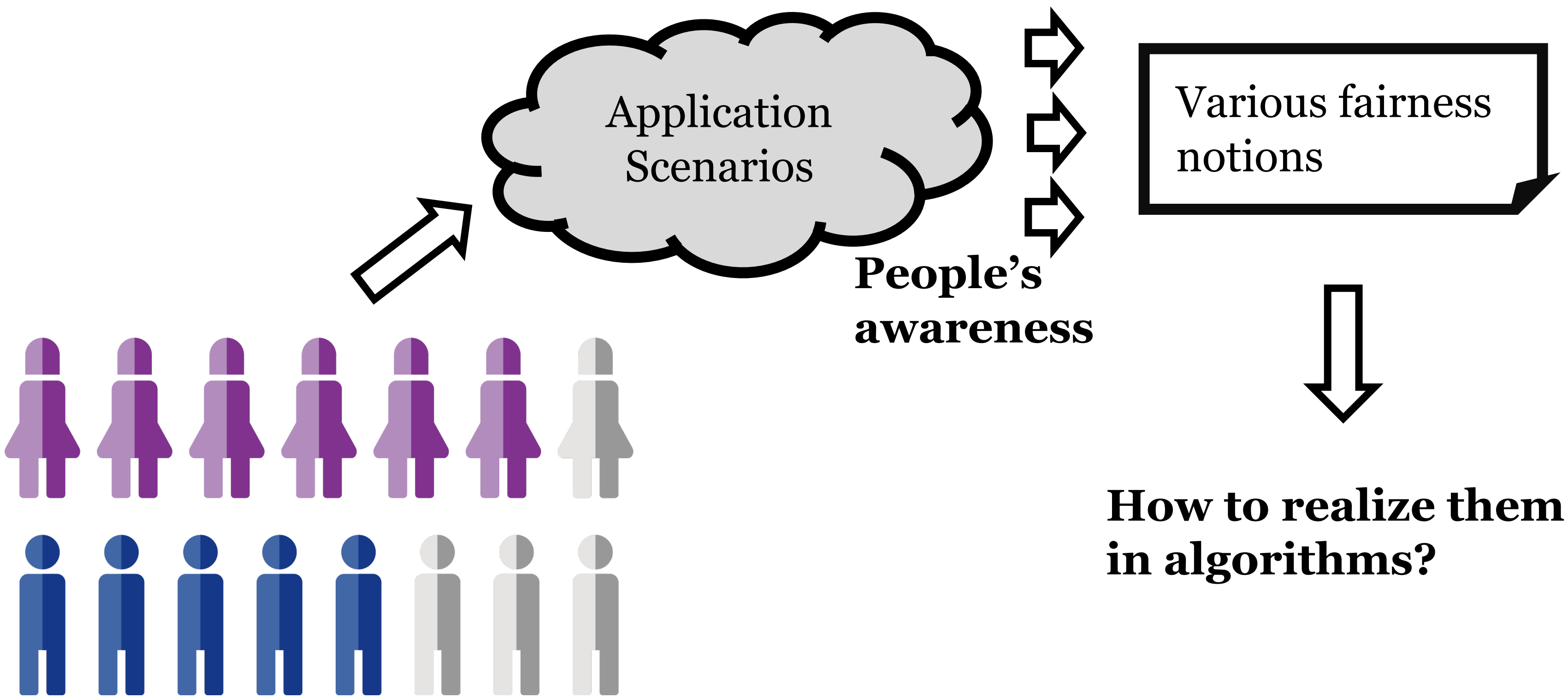
For example, it **depends on the specific studied problem** to determine which case should be considered as fair.

[1] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. IEEE Intelligent Systems, 2020.
[2] Yushun Dong, Jing Ma, Chen Chen, and Jundong Li. Fairness in Graph Mining: A Survey. arXiv preprint arXiv:2204.09888, 2022.

Algorithmic Fairness (Cont.)

Then how to define fairness?

Despite the lack of a **universal criterion** for fairness, we could still study fairness in algorithms: there are **various existing fairness notions** based on people's awareness.



Fairness Notions

- Group fairness
  - The population can be divided into different groups w.r.t. sensitive features (e.g., age, gender, race, …)
  - "no biases towards any specific sensitive group"
- Individual fairness
  - "Similar individuals should have similar prediction results"

Notations

- Tabular Data
  - Sensitive Attribute S (e.g., gender)
  - Non-sensitive Attribute X (e.g., high school grades)
  - Label/ground-truth Y (e.g., university grades)
- Algorithmic Decision-Making
  - Policy/Predictor h predicts label/ground-truth (e.g., graduation) to take decisions (e.g., university admission)

Outline

- Fairness in machine learning
    - Group fairness, individual fairness
    - Causal fairness
- Achieving counterfactual fairness
    - 3-step method
    - Generative models

Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

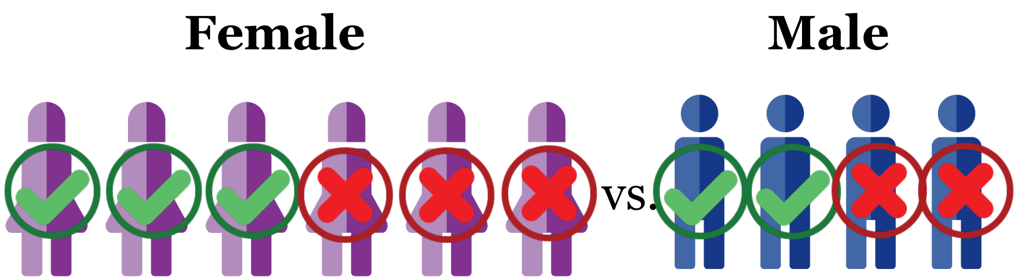
[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.



[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

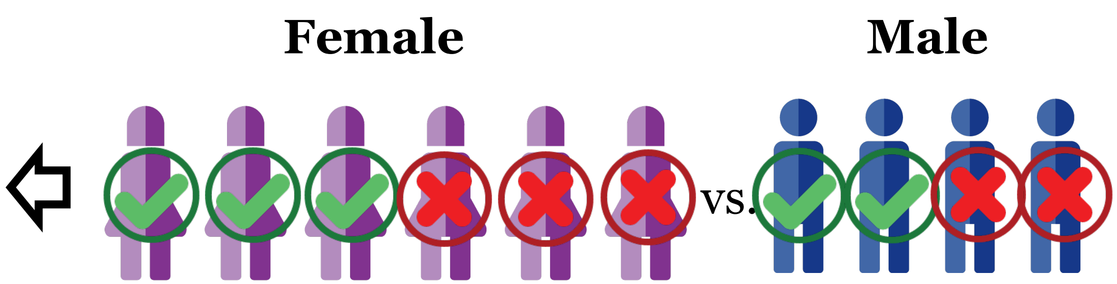
Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Fair in perspective of Demographic Parity.



[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

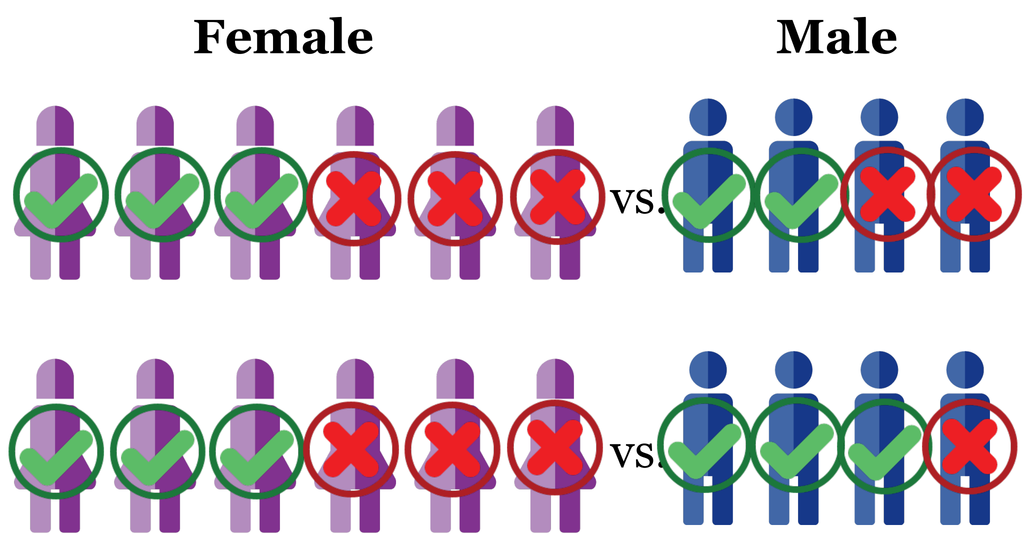
Demographic Parity

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Fair in perspective of Demographic Parity.



[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

Demographic Parity

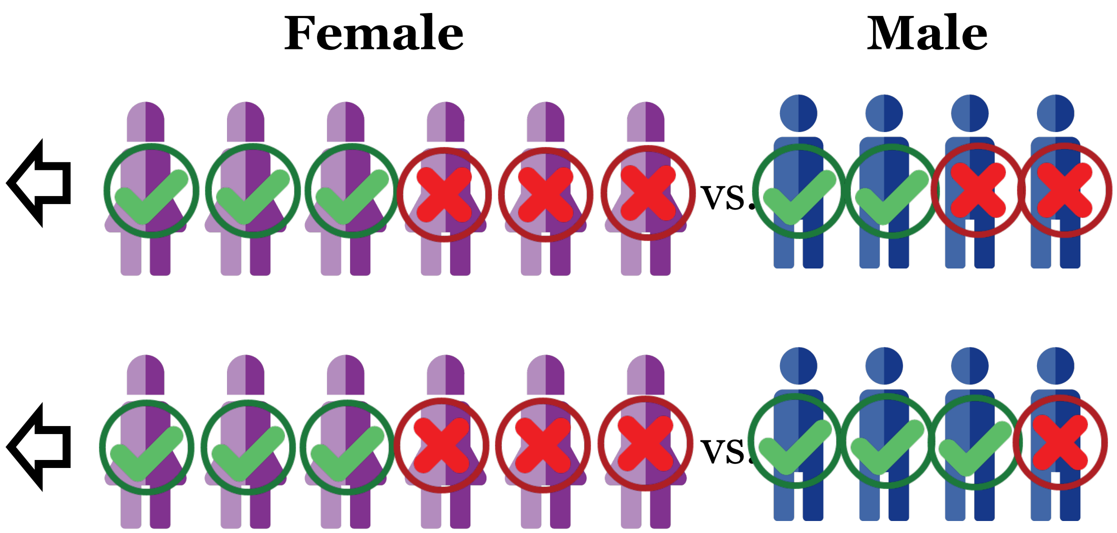
Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Fair in perspective of Demographic Parity.

Unfair in perspective of Demographic Parity.



[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

Demographic Parity (Cont.)

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Criterion: $\mathrm{P}\left(\hat{Y}=1 \middle| S=0\right) = \mathrm{P}\left(\hat{Y}=1 \middle| S=1\right)$

Metric: $\Delta_{DP} = \left|\mathrm{P}\left(\hat{Y}=1 \middle| S=0\right) - \mathrm{P}\left(\hat{Y}=1 \middle| S=1\right)\right|$

[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

Demographic Parity (Cont.)

Group Fairness: Demographic Parity

Demographic Parity is first proposed in **binary classification task** for tabular data [1].

Demographic Parity is considered as achieved if the model yields the **same positive rate** for individuals in both **sensitive subgroups**.

Criterion: $\mathrm{P}\left(\hat{Y}=1 \middle| S=0\right) = \mathrm{P}\left(\hat{Y}=1 \middle| S=1\right)$

Metric: $\Delta_{DP} = \left|\mathrm{P}\left(\hat{Y}=1 \middle| S=0\right) - \mathrm{P}\left(\hat{Y}=1 \middle| S=1\right)\right|$

[1] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Innovations in Theoretical Computer Science, 2012.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

Equality of Odds: the **positive rate** are enforced to be the same between demographic subgroups conditional on the **ground truth class labels**.

[1] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In NeurIPS, 2016.

Equality of Odds/Opportunity

Group Fairness:

Equality of Odds [1] vs. **Equality of Opportunity** [1]

Equality of Odds: the **positive rate** are enforced to be the same between demographic subgroups conditional on the **ground truth class labels**.

Criterion: $\mathrm{P}(\hat{Y} = 1 | S = 0, Y = y) = \mathrm{P}(\hat{Y} = 1 | S = 1, Y = y)$

Metric:

$$\Delta_{EOD} = |P(\hat{Y} = 1 | S = 0, Y = 1) - P(\hat{Y} = 1 | S = 1, Y = 1)| + |P(\hat{Y} = 1 | S = 0, Y = 0) - P(\hat{Y} = 1 | S = 1, Y = 0)|$$

The intuition of Equality of Odds: to enforce the true positive rate (right and positive results) and false positive rate (wrong but positive results) to be the same across groups;

[1] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In NeurIPS, 2016.

Limitation of Statistical Fairness Notions

- Group fairness
  - “no biases towards any specific sensitive group”
  - Difficulty: capturing discrimination without ”causal story”, defining groups
- Individual fairness
  - “Similar individuals should have similar prediction results”
  - Difficulty: defining a similarity function

Berkeley admissions scenario

Men		Women	
Applied	Admitted (%)	Applied	Admitted (%)
8442	44	4321	35

Evidence of discrimination?

[1] Barocas, et al. "Fairness and Machine Learning.", 2019.

Berkeley admissions scenario

	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

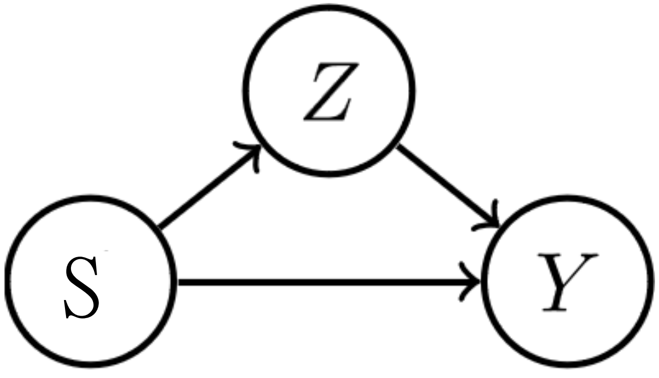
Need to understand the causal mechanism that generated the results we see

Outline

- Fairness in machine learning
  - Group fairness, individual fairness
  - Causal fairness
- Achieving counterfactual fairness
  - 3-step method
  - Generative models

The causal perspective on algorithmic fairness

Causal Graphs: represent causal relationships between variables (nodes of the graph) through the edges of the graph.



S: Gender
Y: Admission
Z: department choice

The causal perspective on fairness

- **Task 1: Discrimination discovery:**
  direct and indirect discrimination, **causal fairness notions**

Part I

- **Task 2: Discrimination removal:**
  learn policies that decide irrespective of sensitive attributes

Part II

Counterfactual Fairness

- A natural question of fairness – What if ?

Will my application get approved if my gender/race/age had been different?

- **Counterfactual fairness**: fairness from a **causal** perspective
  - compare the predictions of each individual from the original data and the counterfactuals with different values of the sensitive attribute

[1] Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness[J]. NeurIPS, 2017.

Counterfactual Fairness

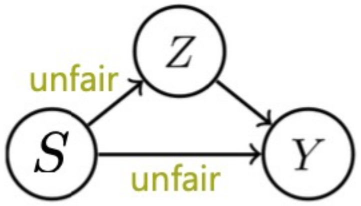
- Had I been assigned sensitive feature S=s', would I have gotten the same decision?

  - **Counterfactual Fairness**

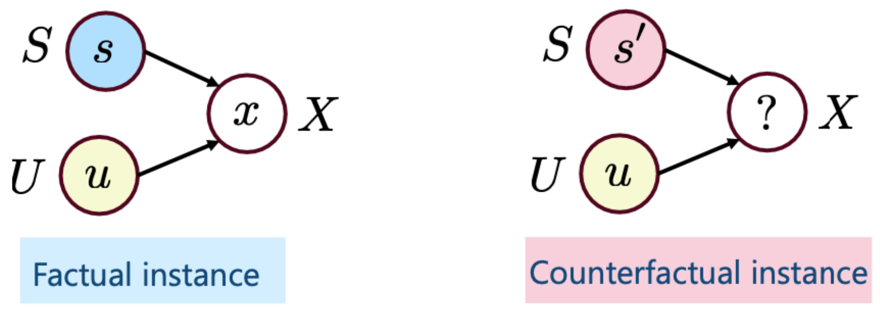
Predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $S = s$,

$$P\left(\hat{Y}_{S \leftarrow s}(U) = y \mid X = x, S = s\right) = P\left(\hat{Y}_{S \leftarrow s'}(U) = y \mid X = x, S = s\right),$$

for all y and for any value s' attainable by S.

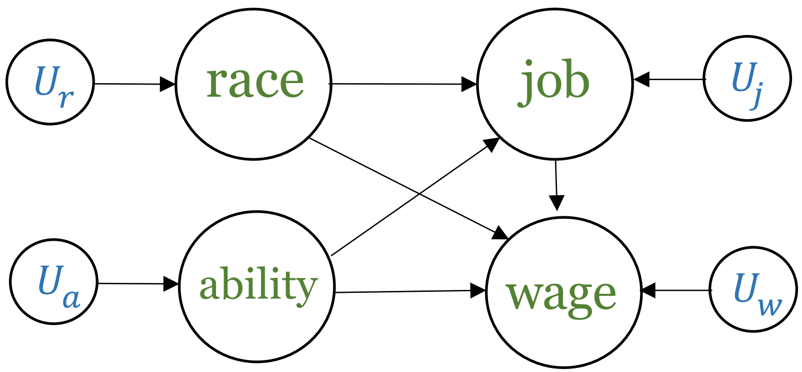


Notice: in counterfactual S←s', other features may change correspondingly.



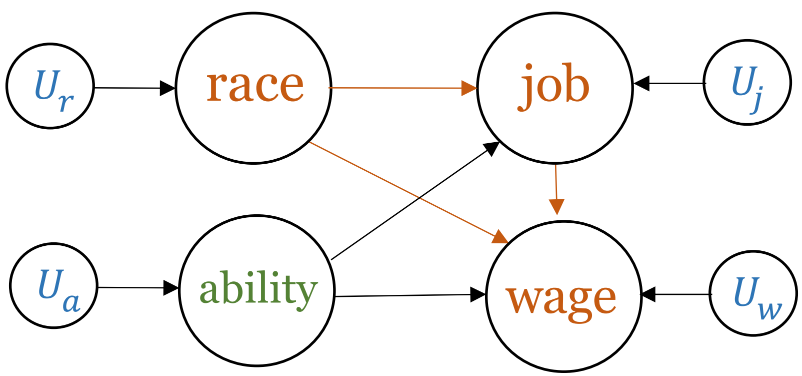
Background: Causal Model

- Structural causal model [Pearl, 2005]
  - Independent exogenous variables (U)
  - Endogenous variables (V)
  - Structural equations (F) (functions which describe the relations between variables)



Background: Causal Model

- Structural causal model [Pearl, 2005]
  - Independent exogenous variables (U)
  - Endogenous variables (V)
  - Structural equations (F) (functions which describe the relations between variables)



Biased information

Difference between interventional and counterfactual fairness

Definition. *A predictor* $\hat{Y}$ *is* **counterfactually fair** *if given observations* $\mathcal{X} = \mathbf{x}$ *and* $A = a$ *we have that,*

$$\mathbb{P}(\hat{Y}_{A \leftarrow a} = y \mid \mathcal{X} = \mathbf{x}, A = a) = \mathbb{P}(\hat{Y}_{A \leftarrow a'} = y \mid \mathcal{X} = \mathbf{x}, A = a)$$

for all y *and* $a' \neq a$.

Compares **the same individual** with a different version of him/herself

Definition (Kilbertus et al., 2017). *A predictor* $\hat{Y}$ *exhibits no* **individual proxy discrimination** *if given observations* $\mathcal{X} = \mathbf{x}$ *and* $A = a$ *we have that,*

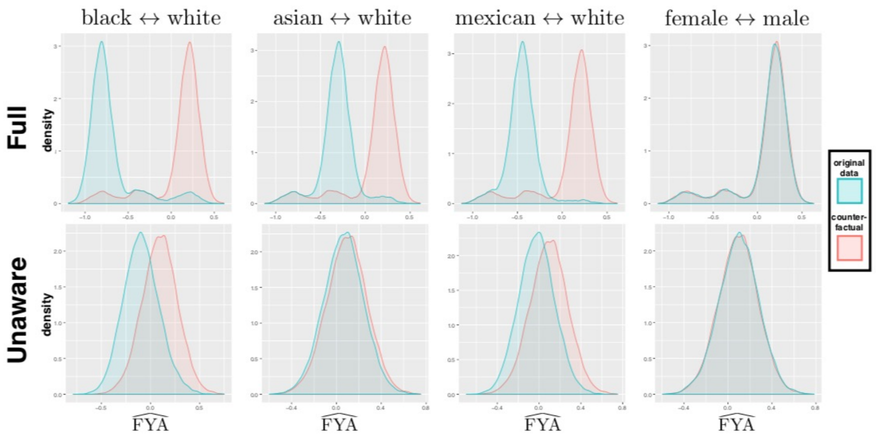
$$\mathbb{P}(\hat{Y} = y \mid do(A = a), \mathcal{X} = \mathbf{x}) = \mathbb{P}(\hat{Y} = y \mid do(A = a'), \mathcal{X} = \mathbf{x})$$

for all y *and* $a' \neq a$.

Compares **different individuals** with the same observed features

How to test the counterfactual fairness of a predictor?

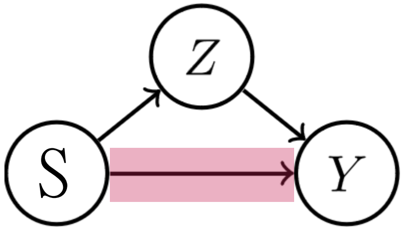
- Empirically test whether the predictors are counterfactually fair
  - 1. Assume the true model of the world is given by a specific causal model
  - 2. Fit the parameters of this model using the observed data
  - 3. Generate samples from the model given either the observed race and sex, or *counterfactual* race and sex variables.
  - 4. Compare the predictions of the classifier to both the original and counterfactual data



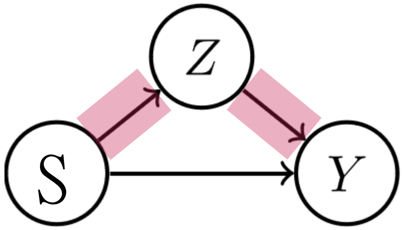
If a predictor is counterfactually fair => the distributions of these two predictions would be similar

Discrimination through different paths

Direct and indirect discrimination



Direct discrimination



Indirect discrimination

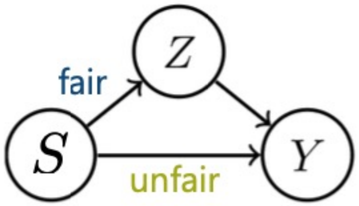
Path-specific Counterfactual Fairness

- Had I been assigned S=s’ - but I keep my today’s department choice, would I have gotten the same decision?

Predictor $\hat{Y}$ is path-specific counterfactually fair if under any context $X = x$ and $S = s$, if it’s prediction coincides with the one that would have been made in a counterfactual world in which along the unfair pathways (denoted by π) $S = s'$:

$$P\left(\hat{Y}_{S\rightarrow s_\pi}(U) = y \mid X, S = s\right) = P\left(\hat{Y}_{S\rightarrow s'_\pi}(U) = y \mid X, S = s\right)$$

for all y and for any value s' attainable by S.

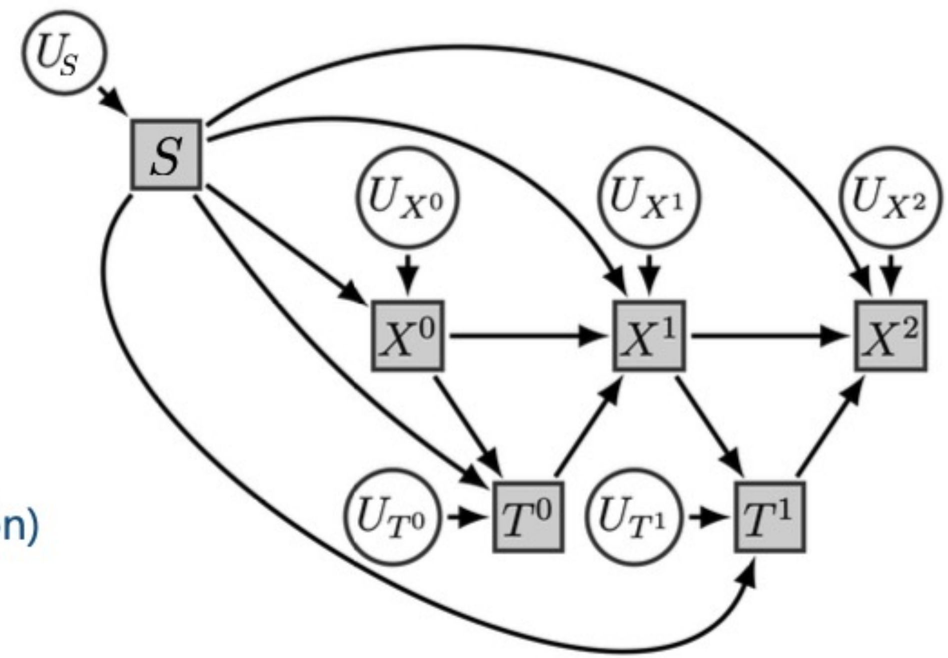


Path specific effect (PSE) = average difference between observed and counterfactual predictions (for given (unfair) paths)

Sylvia Chiappa ”Path- specific counterfactual fairness”, AAAI 2019.

Causal fairness: Long-term perspective

Causal Modeling for Fairness in Dynamical Systems



S: sensitive
(e.g., gender)

X: non-sensitive
(e.g., qualification)

T: treatment
(e.g., decision)

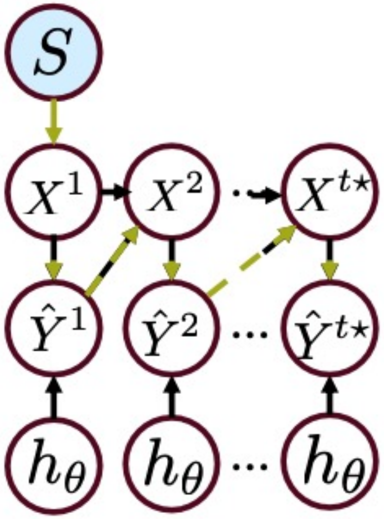
Creager et al., "Causal Modeling for Fairness in Dynamical Systems", ICML 2020.

Causal fairness: Long-term perspective

A predictor h_θ is long-term fair, if

$$P\left(\hat{Y}_{S\rightarrow s_\pi,\theta}(U) = y \mid X = x, S = s\right) = P\left(\hat{Y}_{S\rightarrow s'_\pi,\theta}(U) = y \mid X, S = s\right)$$

where π is a set of paths from S to $\hat{Y}^{t\star}$ via $\{X^n, \hat{Y}^n\}_{n=1}^{t\star}$
and θ is a soft intervention through all paths.



S: sensitive X: non-sensitive. Y: Prediction/Decision. h: Policy model

The causal perspective on fairness

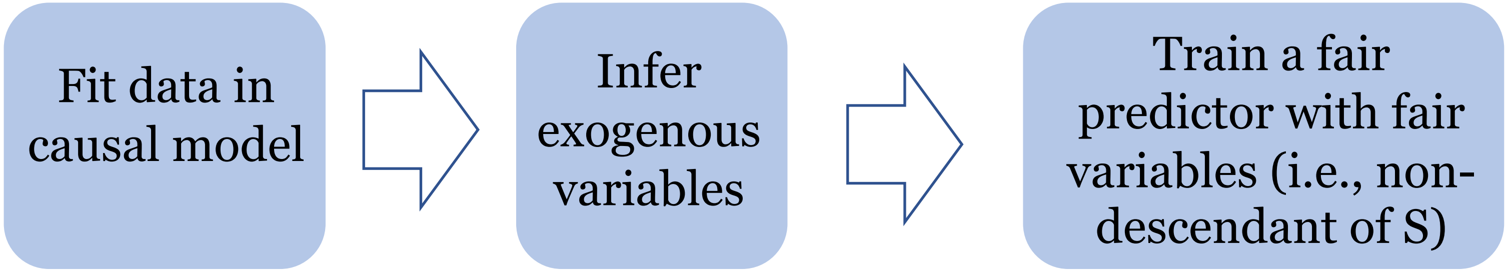
- **Task 1: Discrimination discovery:** Part I
  direct and indirect discrimination, **causal fairness notions**
- **Task 2: Discrimination removal:** Part II
  learn policies that decide irrespective of sensitive attributes

Outline

- Fairness in machine learning
  - Group fairness, individual fairness
  - Causal fairness
- Achieving counterfactual fairness
  - 3-step method
  - Generative models

How do we implement counterfactual fairness?

- Depends on context, prior causal knowledge and assumptions.
- If we have (partial) prior knowledge of causal model:



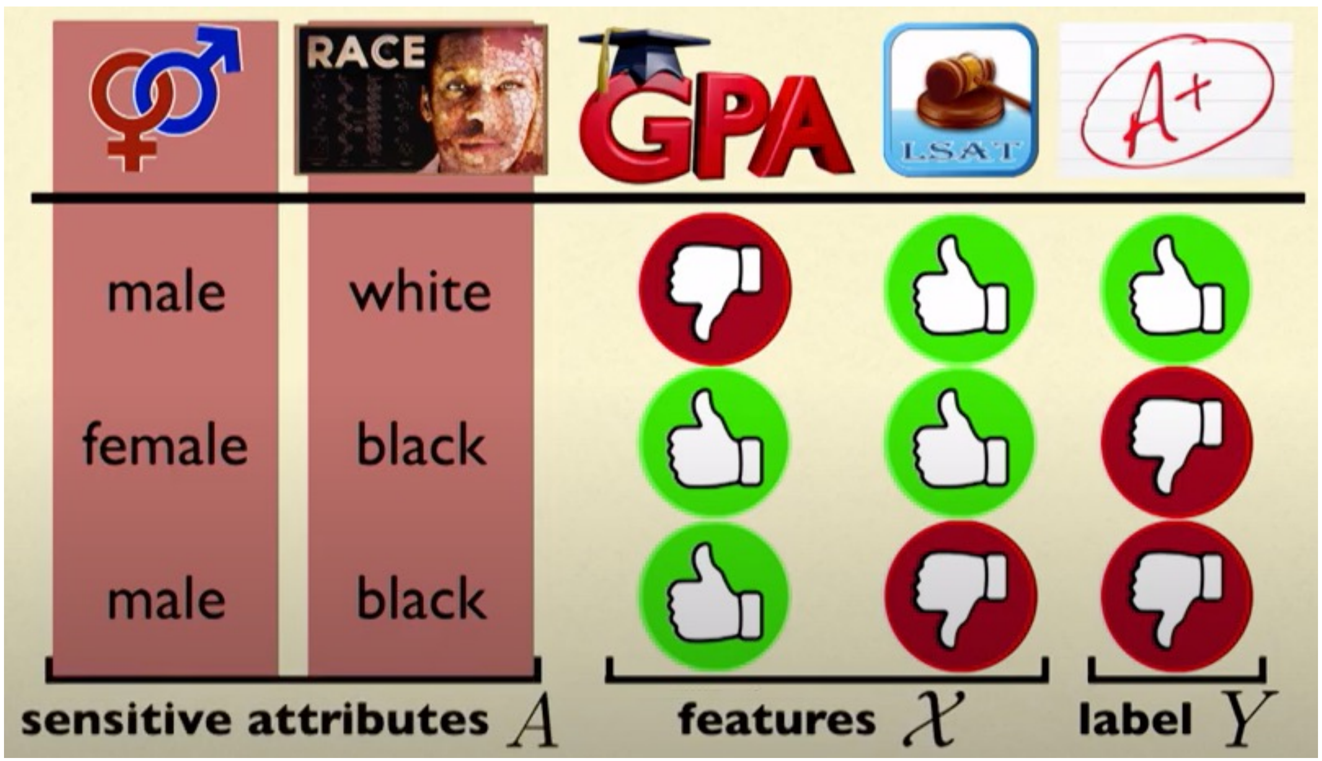
Three Steps in Computing Counterfactuals

- Counterfactual inference, as specified by a causal model (U, V, F) given evidence W, is the computation of probabilities $P(Y_{Z \leftarrow z}(U) | W = w)$, where W, Z and Y are subsets of V. Inference proceeds in three steps:

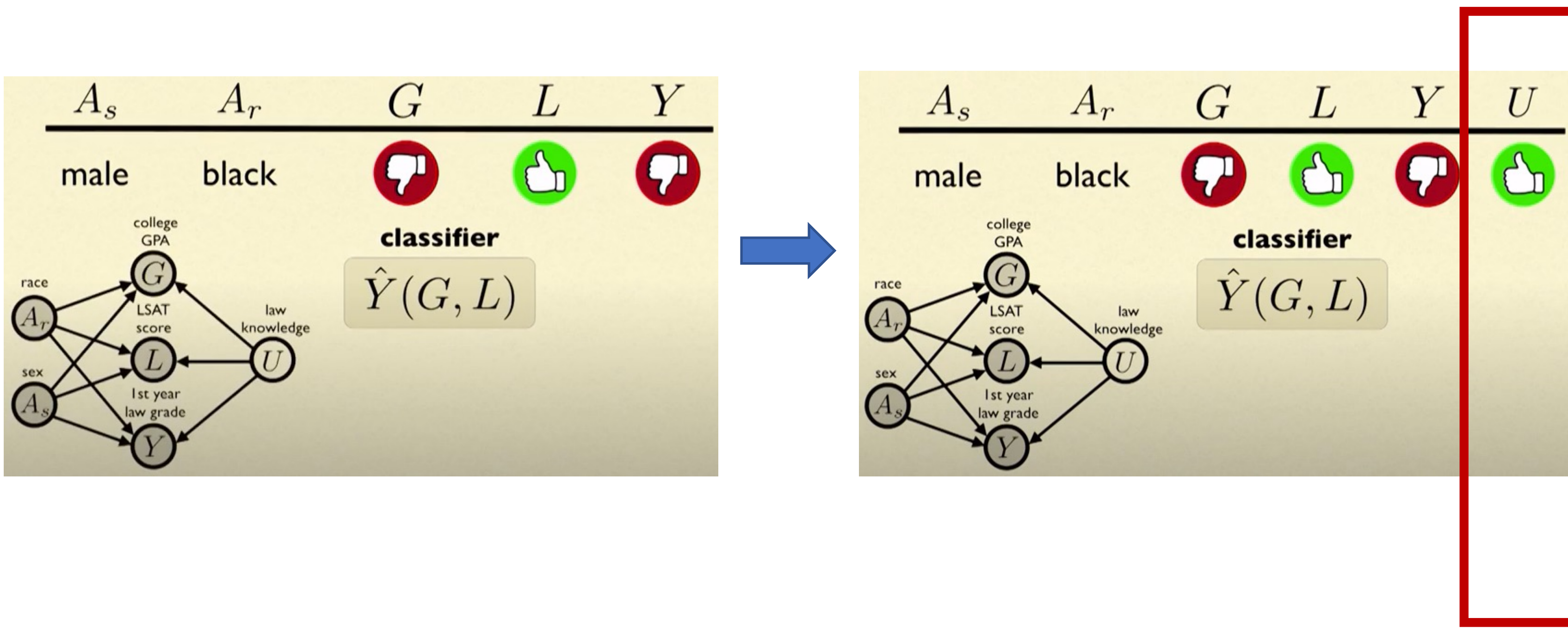
- Step 1: **Abduction**: for a given prior on U, compute the posterior distribution of U given the evidence W = w
- Step 2: **Action**: Modify the model, M, by removing the structural equations for the variables in Z and replacing them with the appropriate functions Z = z, resulting in the modified set of equations F_z;
- Step 3: **Prediction**: compute the implied distribution on the remaining elements of V using F_z and the posterior $P(U \mid W = w)$.

Example: Law school

- Aim: Predict students' first-year average grade (FYA)
- Sensitive features: race, gender

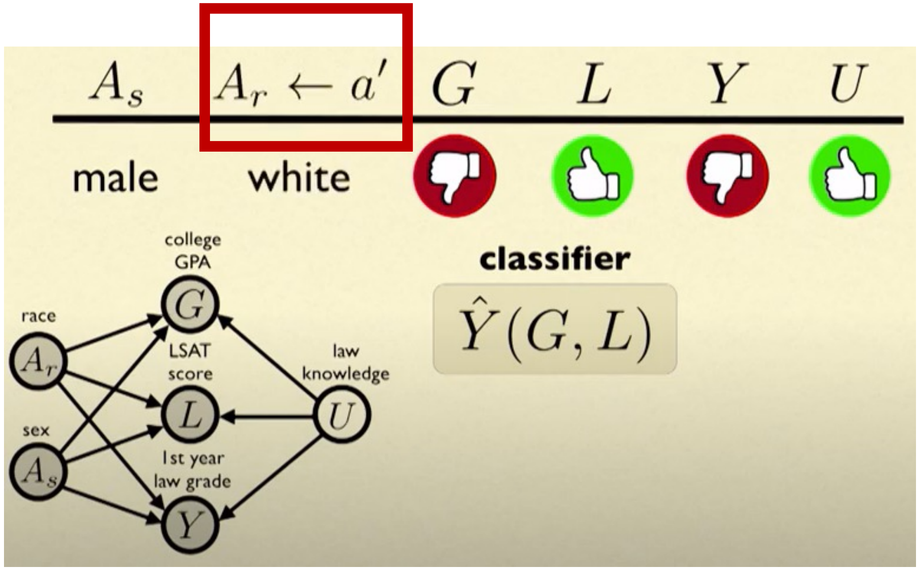


Step 1: Compute exogenous variables in the causal model



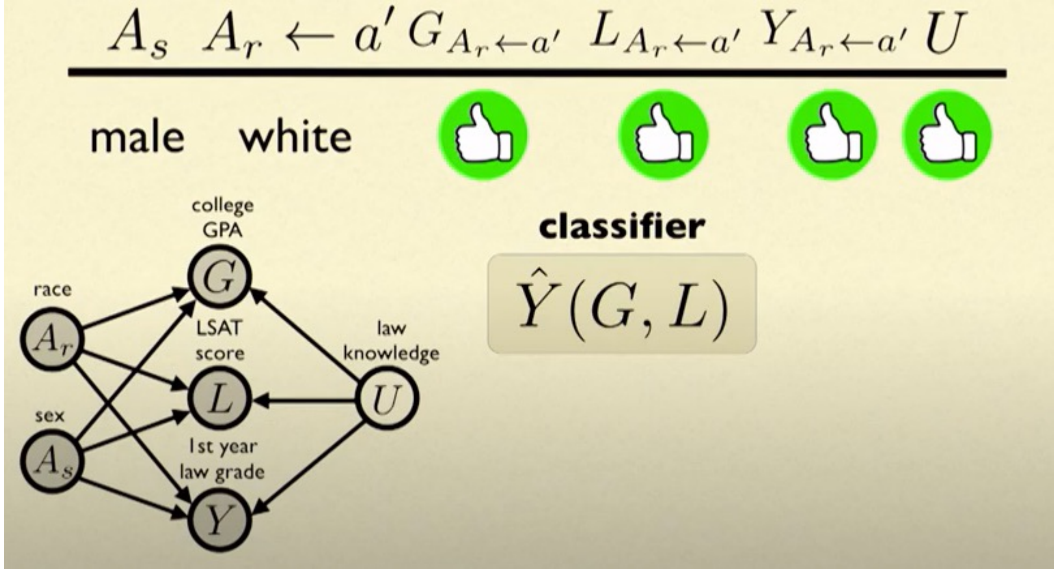
Exogenous variables capture feature information without bias of S

Step 2: Change protected attributes



3. Recompute observed variables in the causal model

- Get counterfactual results



A counterfactual fairness algorithm

Given: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}, a^{(i)})\}_{i=1}^{d}$

a) Fit causal model $\mathcal{M}$

b) For each data point $i \in \mathcal{D}$, compute $u^{(i)}$

c) $\hat{\theta} \leftarrow \arg\min_{\theta} \sum_{i \in \mathcal{D}} \ell(y^{(i)}, \hat{Y}_{\theta}(u^{(i)}, \mathbf{x}_{\not\succ A}^{(i)}))$

d) Return: $\hat{Y}_{\hat{\theta}}$

features that are
not descendants of A

How do we implement counterfactual fairness?

- Depends on context, prior causal knowledge and assumptions.
  - If we have (partial) prior knowledge of causal model
  - If we have no parametric causal knowledge
    - Probabilistic approximation → deep generative models

Outline

- Fairness in machine learning
  - Group fairness, individual fairness
  - Causal fairness
- Achieving counterfactual fairness
  - 3-step method
  - Generative models

Why do we need generative models?

Nature of structural equations and **relations unknown**

- **Causal graph is known**

Nature of individual **exogenous** factors **unknown**

- **Consider some hidden latent factors exist**

Need to **estimate** causal **functions** and **exogenous** factors!

Why do we need generative models?

Special class of **neural network** models that use unlabeled data to estimate unknown data distribution.

Two broad, popular types of generative models:

1. Variational autoencoders (**VAE**)
2. Generative adversarial networks (**GAN**)

Why do we need generative models?

Special class of **neural network** models that use unlabeled data to estimate unknown data distribution.

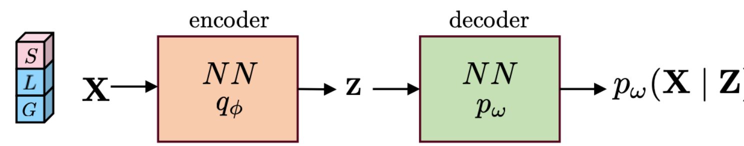
Two broad, popular types of generative models:

Our focus today

1. Variational autoencoders (**VAE**)
2. Generative adversarial networks (**GAN**)

Variational autoencoders: A (very) brief overview

Estimate data distribution through lower dimensional latent space.



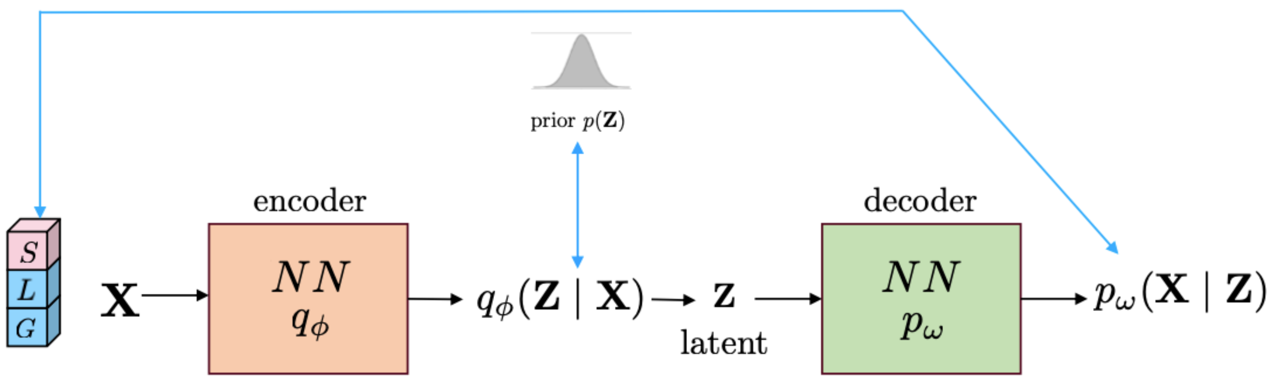
Two neural networks learnt jointly:

- **Encoder** learns parameters of latent distribution from data
- **Decoder** learns to use latent factors to regenerate observed data

How does VAE use variational inference to learn data distributions?

Kingma and Welling "Auto-encoding variational Bayes" ICLR 2014.

VAE decoder regenerates original data from latent



$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}\left[-\log p_\omega(\mathbf{X} \mid \mathbf{Z})\right] + \mathrm{KL}\left[q_\phi(\mathbf{Z} \mid \mathbf{X}) \,||\, p(\mathbf{Z})\right]$$

Maximize log-likelihood of generated data **Minimize latent space divergence to prior (Gaussian)**

Kingma and Welling "Auto-encoding variational Bayes" ICLR 2014.

How can we use VAEs for causal fairness?

Nature of structural functions and relations unknown

- **Use causal graph in VAEs to estimate structural relations**

Nature of individual exogenous factors unknown

- **Encode unbiased causal latent factors from data using VAE**

How can we train counterfactually fair predictors?

How can we use VAEs for causal fairness?

Nature of structural functions and relations unknown

- **Use causal graph in VAEs to estimate structural relations**

Nature of individual exogenous factors unknown

- **Encode unbiased causal latent factors from data using VAE**

How can we train counterfactually fair predictors?

Estimate causal relations with VAE models

Causal estimation: Variational Graph Autoencoders

Estimate data distribution using causal graph with **VACA**.

- **Input causal** graph information in the model *(adjacency matrix)*
- Latent factors **capture exogenous information** for each feature

How can variational inference **learn causal relations**?

Sánchez-Martin et al. “VACA: Designing Variational Graph Autoencoders for Causal Queries”, AAAI 2022.

Causal estimation: Variational Graph Autoencoders

Estimate data distribution using causal graph with **VACA**.

- **Input causal** graph information in the model *(adjacency matrix)*
- Latent factors **capture exogenous information** for each feature

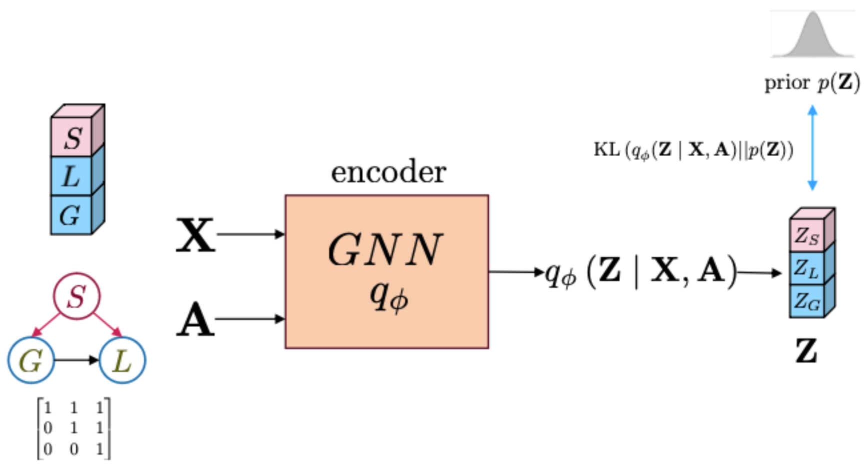
Encoder, Decoder are **Graph Neural Networks** [Kipf and Welling, 2017]

- Help learn (direct and indirect) causal relationships
- Estimates structural causal functions between features

How can variational inference **learn causal relations**?

Sánchez-Martin et al. “VACA: Designing Variational Graph Autoencoders for Causal Queries”, AAAI 2022.

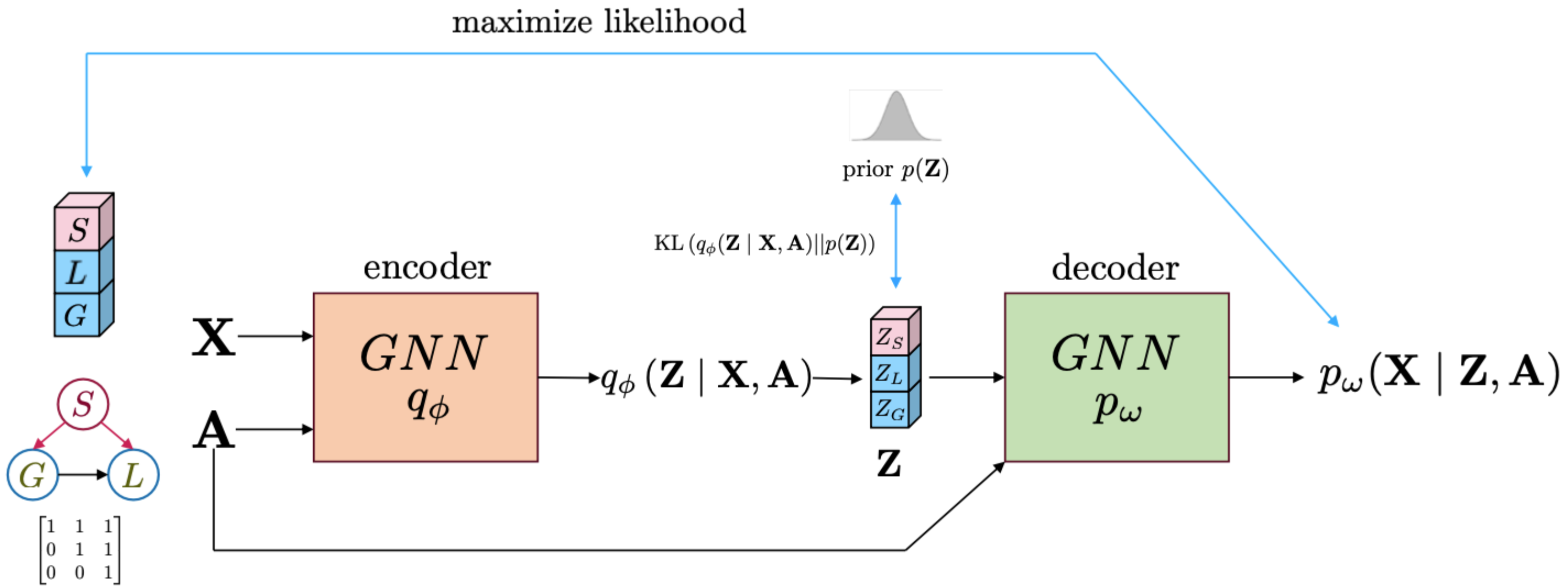
Estimating causal relations with VACA



Latent factors capture **information** of **each feature independent** of **parents**' effect

Sánchez-Martin et al. “VACA: Designing Variational Graph Autoencoders for Causal Queries”, AAAI 2022.

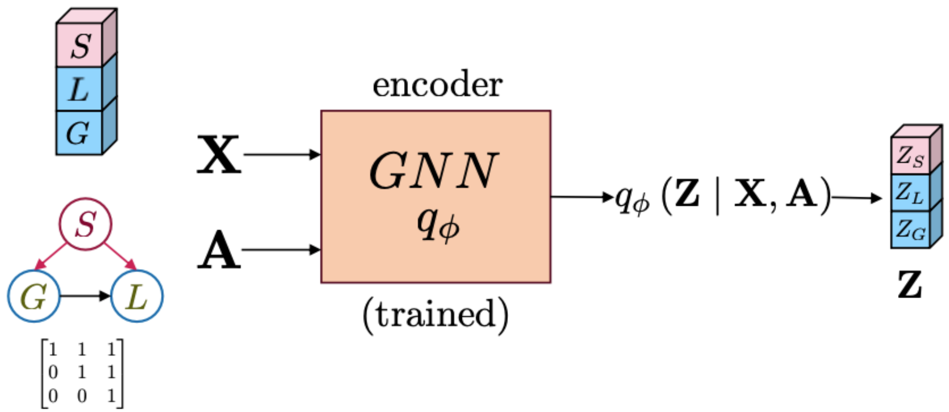
Estimating causal relations with VACA



How can we use VACA for **counterfactual fairness**?

Sánchez-Martin et al. “VACA: Designing Variational Graph Autoencoders for Causal Queries”, AAAI 2022.

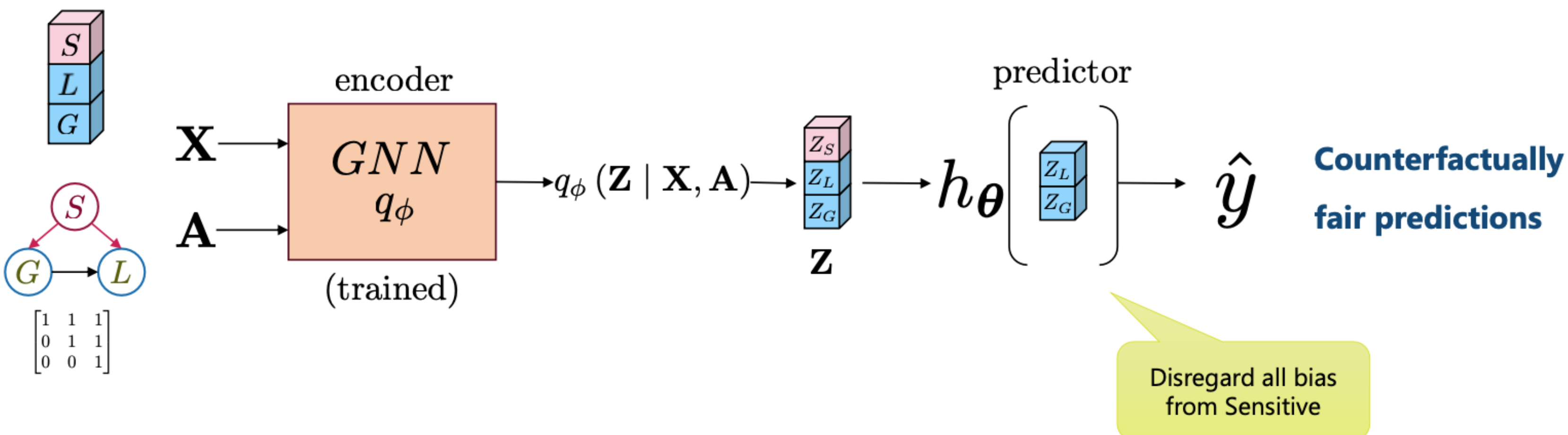
Counterfactually fair prediction with trained VACA



Latent factors capture **information** of **each feature independent** of **parents**' effect

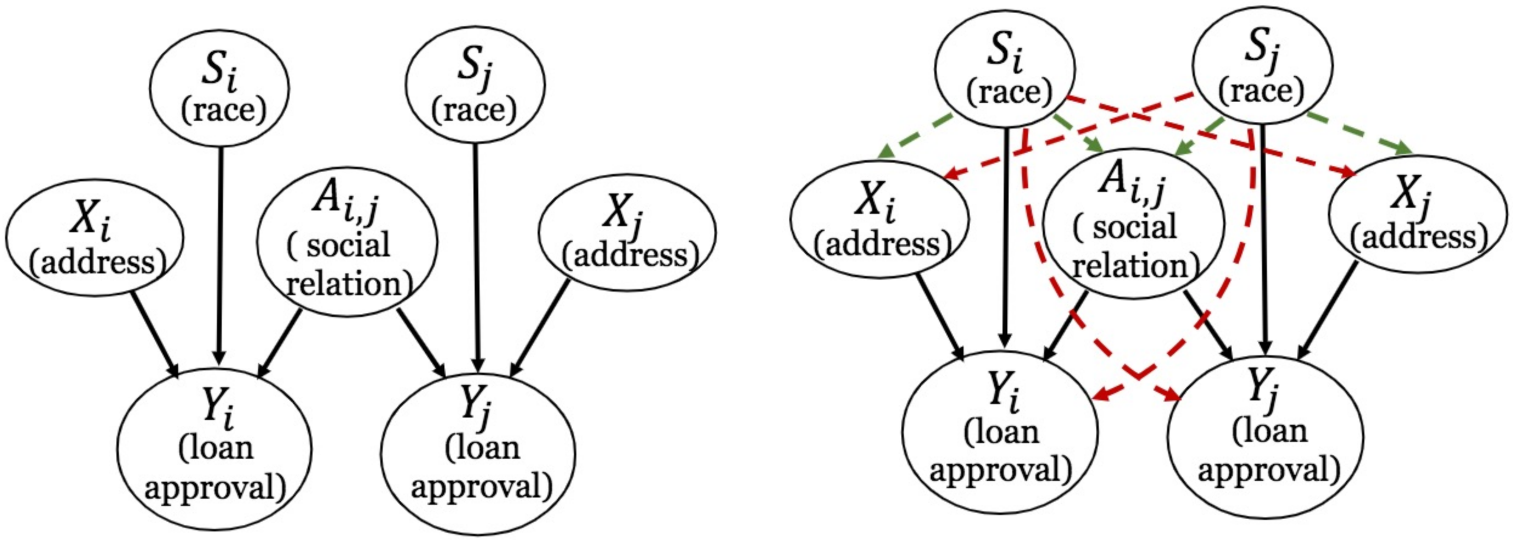
Sánchez-Martin et al. "VACA: Designing Variational Graph Autoencoders for Causal Queries", AAAI 2022.

Counterfactually fair prediction with trained VACA



Sánchez-Martin et al. “VACA: Designing Variational Graph Autoencoders for Causal Queries”, AAAI 2022.

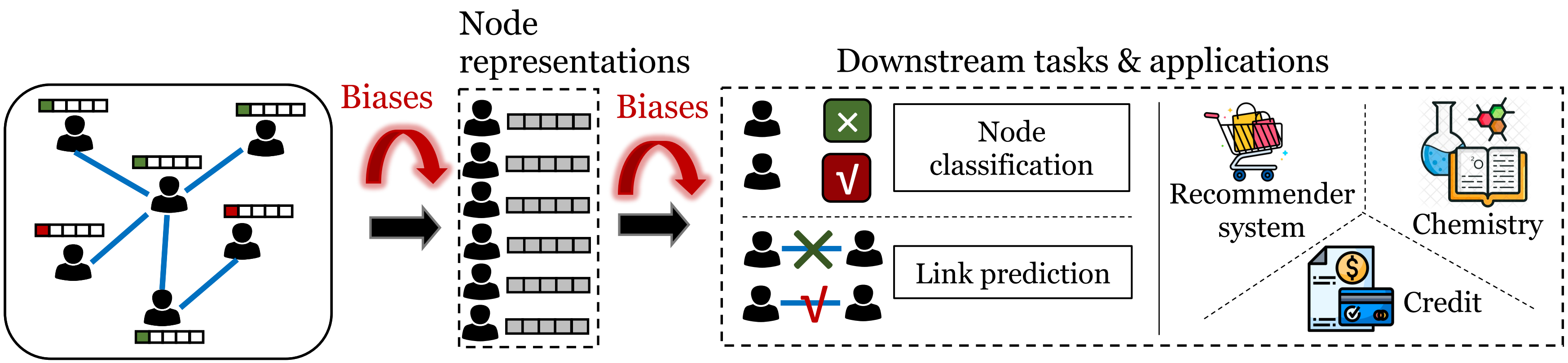
Counterfactual Fairness on Graphs



- Limitations of the above fairness notion:
  - In graphs, the sensitive attributes of each node's neighbors may causally affect the prediction w.r.t. this node (red dashed edges);
  - The sensitive attributes may causally affect other features and the graph structure (green dashed edges).

Bias in Node Representations

- **Node representation learning**: map nodes into a latent embedding space to facilitate various downstream tasks such as node classification



- Limitation: the representations may contain biases towards certain sensitive attribute, e.g., race/gender

Graph Counterfactual Fairness

- **Graph counterfactual fairness**: An encoder $Z_i = (\Phi(X, A))_i$ satisfies graph counterfactual fairness if for any node i:

$$P(\boxed{(Z_i)_{S \leftarrow \mathbf{s}'}} | X = \mathbf{X}, A = \mathbf{A}) = P((Z_i)_{\boxed{S} \leftarrow \mathbf{s}''} | \boxed{X = \mathbf{X},} \boxed{A = \mathbf{A}}),$$

Node representation for node i after intervention on S with value s'

Sensitive features

Node features

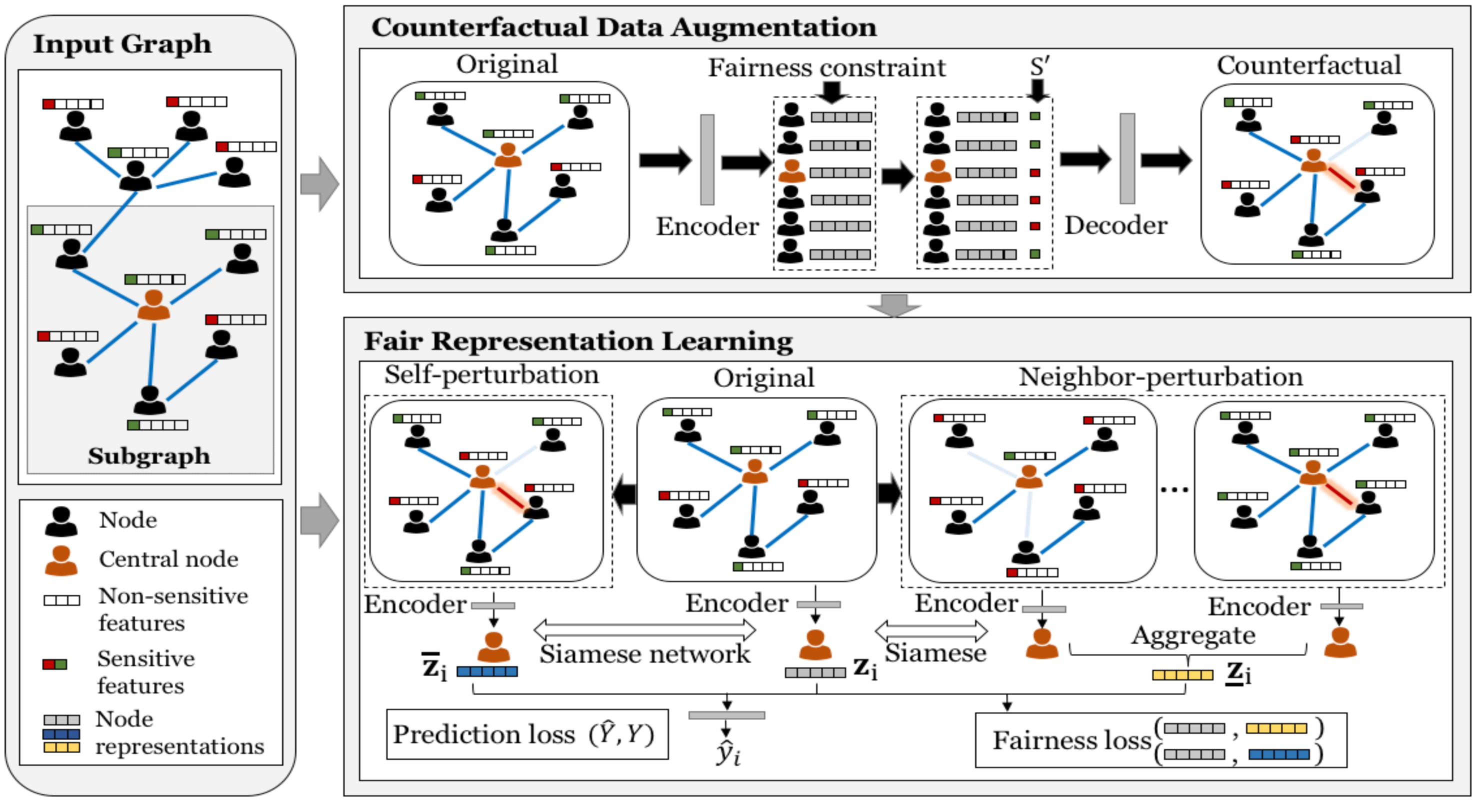
Graph structure

- Example: the prediction for one's loan application should be the same regardless this applicant's <u>and his/her friends' (connected in a social network)</u> race information

Ma J, Guo R, Wan M, et al. Learning fair node representations with graph counterfactual fairness. WSDM. 2022: 695-703.

GEAR

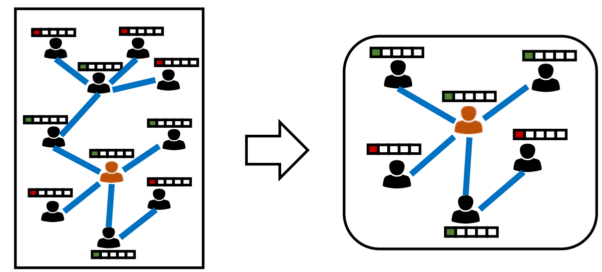
Aim: learn node representations on graph towards graph counterfactual fairness, and maintain a good prediction performance simultaneously



Ma J, Guo R, Wan M, et al. Learning fair node representations with graph counterfactual fairness. WSDM. 2022: 695-703.

Proposed Framework

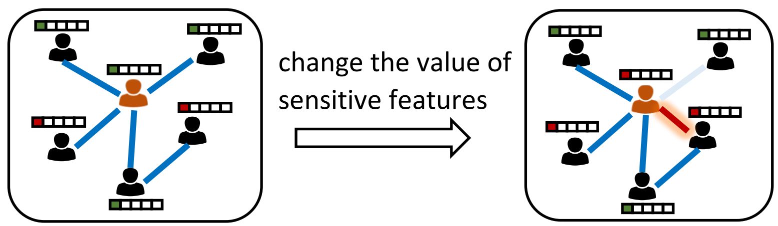
- **Subgraph generation**: the graph is often very large. We split it into small subgraphs for each centroid node for better efficiency



Graph Subgraph

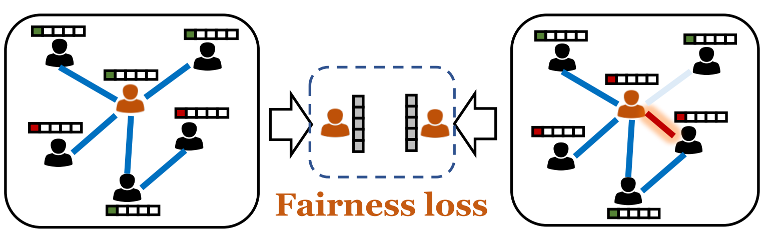
- **Counterfactual (CF) augmentation**: generate CFs for each subgraph with perturbation on sensitive features of different nodes

Original



Counterfactual

- **Fair representation learning**: learn fair representations which elicit the same predicted label across different CFs w.r.t. the same node



Original Counterfactual

$$\mathcal{L}_f = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \left((1-\lambda_s) \boxed{d(\mathbf{z}_i, \overline{\mathbf{z}}_i)} + \lambda_s \boxed{d(\mathbf{z}_i, \underline{\mathbf{z}}_i)} \right)$$

Fairness loss: Encourage the node representations learned from the original graph and CFs to be the same

Ma J, Guo R, Wan M, et al. Learning fair node representations with graph counterfactual fairness. WSDM. 2022: 695-703.

Evaluation

Observations:

- GEAR achieves comparable prediction performance as state-of-the-art node representation learning methods
- GEAR performs well in different fairness notions, especially outperforms all baselines in graph counterfactual fairness

Metric for graph counterfactual fairness

Dataset	Method	Prediction Performance			Fairness			
		Accuracy (↑)	F1-score (↑)	AUROC (↑)	Δ_{EO} (↓)	Δ_{DP} (↓)	δ_{CF} (↓)	R^2 (↓)
Synthetic	GCN	0.686 ± 0.015	0.687 ± 0.020	0.758 ± 0.017	0.050 ± 0.030	0.060 ± 0.033	0.101 ± 0.030	0.085 ± 0.050
GNN	GraphSAGE	0.712 ± 0.012	0.714 ± 0.021	0.789 ± 0.018	0.049 ± 0.036	0.053 ± 0.042	0.172 ± 0.056	0.011 ± 0.011
	GIN	0.682 ± 0.021	0.691 ± 0.022	0.741 ± 0.021	0.077 ± 0.053	0.081 ± 0.055	0.301 ± 0.080	0.011 ± 0.009
	C-ENC	0.665 ± 0.023	0.671 ± 0.031	0.732 ± 0.028	0.030 ± 0.024	0.048 ± 0.026	0.633 ± 0.013	0.085 ± 0.016
GNN+Fairness	FairGNN	0.668 ± 0.020	0.672 ± 0.026	0.735 ± 0.022	**0.025 ± 0.021**	**0.042 ± 0.033**	0.678 ± 0.014	0.091 ± 0.021
	NIFTY-GCN	0.618 ± 0.035	0.640 ± 0.037	0.672 ± 0.042	0.172 ± 0.110	0.199 ± 0.106	0.208 ± 0.090	0.105 ± 0.081
	NIFTY-SAGE	0.664 ± 0.041	0.682 ± 0.073	0.755 ± 0.021	0.031 ± 0.027	0.048 ± 0.027	0.147 ± 0.071	0.008 ± 0.005
Our method	GEAR	**0.718 ± 0.018**	**0.724 ± 0.022**	**0.793 ± 0.014**	0.052 ± 0.038	0.064 ± 0.038	**0.002 ± 0.002**	**0.007 ± 0.006**
Bail	GCN	0.838 ± 0.017	0.782 ± 0.023	0.885 ± 0.018	0.023 ± 0.019	0.075 ± 0.014	0.132 ± 0.059	0.075 ± 0.028
	GraphSAGE	**0.854 ± 0.026**	**0.804 ± 0.032**	**0.905 ± 0.021**	0.039 ± 0.022	0.086 ± 0.039	0.088 ± 0.047	0.069 ± 0.011
	GIN	0.731 ± 0.058	0.656 ± 0.084	0.773 ± 0.069	0.041 ± 0.023	0.065 ± 0.034	0.143 ± 0.069	0.047 ± 0.036
	C-ENC	0.842 ± 0.047	0.792 ± 0.014	0.889 ± 0.033	0.038 ± 0.022	0.069 ± 0.020	0.040 ± 0.025	0.078 ± 0.024
	FairGNN	0.835 ± 0.024	0.784 ± 0.021	0.882 ± 0.035	0.046 ± 0.013	0.074 ± 0.026	0.042 ± 0.032	0.086 ± 0.016
	NIFTY-GCN	0.752 ± 0.065	0.669 ± 0.050	0.799 ± 0.051	0.019 ± 0.015	**0.036 ± 0.022**	0.031 ± 0.017	0.025 ± 0.018
	NIFTY-SAGE	0.823 ± 0.048	0.723 ± 0.103	0.876 ± 0.043	**0.014 ± 0.006**	0.047 ± 0.015	0.013 ± 0.011	0.044 ± 0.020
	GEAR	0.852 ± 0.026	0.800 ± 0.031	0.896 ± 0.016	0.019 ± 0.023	0.058 ± 0.017	**0.003 ± 0.002**	**0.038 ± 0.012**

Ma J, Guo R, Wan M, et al. Learning fair node representations with graph counterfactual fairness. WSDM. 2022: 695-703.

Other methods for causal fairness in predictions

- **Fair data generation:** Use causal knowledge in GANs, generate counterfactual fair data for training [Xu et al., 2019]
- **Post-processing**: Given trained classifier, modify outputs to gain counterfactual fairness [Wu et al., 2019]
- **Adversarial learning**: Enforce counterfactual fairness through adversarial constraints [Grari et al., 2023]
- **Path-specific fairness**: VAE based model to satisfy path-specific counterfactual fairness [Chiappa, 2019]

Takeaway on bringing causal fairness to practice

- Operationalizing counterfactual fairness depends on what causal estimation is required.
    - Depends on **context**, prior **causal knowledge** and assumptions.

1. Use **simpler methods** (regression) for **exact** computation
    - We **know what the causal relations are** (additive noise)
2. Use more **complex models** (generative) for **approximation**
    - Parametric causal **relations** are **not known**

Causality and fairness go beyond predictions

Causality and fairness can be studied for other aspects of decision-making algorithmic systems

- Causality and the notion of fairness in algorithmic recourse
  - Are sensitive features involved in recourse? How do they cost for recourse?
- Causality and the notion of harm in algorithmic decisions
  - Do particular actions made by an algorithm instigate harm on people?

Causality, Fairness and Recourse

How do we know if **protected** features **cause recourse cost**?

- Statistical studies can only show recourse cost varies across sensitive groups

Causality, Fairness and Recourse

How do we know if **protected** features **cause recourse cost**?

- Statistical studies can only show recourse cost varies across sensitive groups

Causal knowledge helps us out

- Find sensitive counterfactuals using causal knowledge
- Recourse action costs for individual vs their sensitive counterfactual

One's **sensitive attribute should not cause** how much **cost** is needed for recourse!

von Kügelgen et al. "On the fairness of causal algorithmic recourse", AAAI 2022.

Causality and Harm

How can we account for **harm** of algorithmic **decisions**?

Social sciences account for harm using:

- **Counterfactual Comparative Account**: Action causes harm if affected person would have been better off without it.

Causal knowledge can help us out

- Did **action cause harm**? How much more utility would we get if the action was not performed?
- **Counterfactual harm: Harm caused** by action given some context of features and outcome, compared to a default action and outcome

Richens et al. "Counterfactual Harm", NeurIPS 2022.

Considerations for implementing causal fairness

We must take care while applying causal methods for algorithmic fairness in practice

- Can we reliably estimate fairness for our specific scenarios?
- Can we ensure fair algorithms if we do not correctly know all the cause-effect relations?
- Can we even use causality in societal settings of fairness?

Counterfactual identifiability and confounding

Causal fairness **critically relies** on **causal specifications.**

1. **Identifiability of counterfactuals**
   - Can we reliably compute counterfactuals from just observed data?
2. **Unmeasured confounding**
   - Are counterfactuals estimates robust to unobserved variations?

Can we reliably estimate counterfactual fairness

when assumptions do not hold?

Counterfactual identifiability and confounding

We can **bound counterfactual fairness estimates**!

1. **Unidentifiability**
   - **Causal graph factorization**[1] can exactly show source of unidentifiability and give estimate bounds

[1] Wu et al. "Counterfactual fairness: Unidentification, bound and algorithm." IJCAI 2019.

Counterfactual identifiability and confounding

We can **bound counterfactual fairness estimates**!

1. **Unidentifiability**
   - **Causal graph factorization**[1] can exactly show source of unidentifiability and give estimate bounds

2. **Confounding**
   - **Algorithms**[2] do **sensitivity analysis** to compute bounds on counterfactual unfairness given assumptions and confounding

Even when causal assumptions are imperfect, we can provide uncertainty bounds for reliability!

[1] Wu et al. "Counterfactual fairness: Unidentification, bound and algorithm." IJCAI 2019.
[2] Kilbertus et al., "The sensitivity of counterfactual fairness to unmeasured confounding

Fairness for partially unknown causal graphs

Most causal fairness works consider full causal graph knowledge.

→ Assume **descendants** and **non-descendants** of **sensitive** are **known**

But in the real-world all causal relations may not be known!

→ Access to **partial causal graph**

Can we get counterfactual fairness with partially unknown causal graph?

Fairness for partially unknown causal graphs

Counterfactual fairness possible with **partial DAGs**

- Some edges undirected, **all cause-effect relations not known**

Ensuring counterfactual fairness with partial knowledge:

- **Identify ancestry** of all features w.r.t. **Sensitive** (definite non-descendants, definite descendants, possible descendants)
- Build **counterfactually fair** predictor using identified **ancestral relations**

Can **exactly identify** relations and get **fairness** if **sensitive** attribute is a **root** node!

Zuo et al. "Counterfactual fairness with partially known causal graph" NeurIPS 2022.

Fairness and causality in the real-world

- Observational analysis is limited. Can only provide minimal information.
- Causality provides a strong foundation for fairness and responsible algorithmic methods
  - Incorporate domain knowledge about societal settings
  - Allows for better mechanism design to incorporate fairness goals
  - Analyze the cause of discrimination in societal settings
  - Framework to study other ethical aspects, e.g., harm
- Starting point for further research into fair and socially responsible systems

References

- Kusner M J, Loftus J, Russell C, et al. Counterfactual fairness[J]. Advances in neural information processing systems, 2017, 30.
- Chiappa S. Path-specific counterfactual fairness[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 7801-7808.
- Wu Y, Zhang L, Wu X. Counterfactual fairness: Unidentification, bound and algorithm[C]//Proceedings of the twenty-eighth international joint conference on Artificial Intelligence. 2019.
- Sánchez-Martin et al. "VACA: Designing Variational Graph Autoencoders for Causal Queries", AAAI 2022.
- Ma J, Guo R, Wan M, et al. Learning fair node representations with graph counterfactual fairness. WSDM. 2022: 695-703.
- Some content of slides are from:
  - Socially Responsible Machine Learning: A Causal Perspective. KDD 2023 Tutorial.

Thank you!
Questions?