TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

"Bias & Fairness"

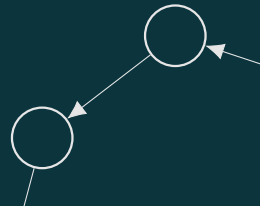
Prof. Dr. Kristian Kersting

Moritz Willig — Today's speaker

Tim Woydt

Florian Busch

Matej Zečević



Course Overview

Pearl. Caus.

1. Introduction
2. Probabilities and Bayesian Networks
3. Structural Causal Models
4. do-Calculus
5. Causal Discovery
6. Uncertainty

Repr. & ML

7. Causal Abstractions
8. Neuro-Causal Models
9. Causal Representation Learning
10. Causality and LLMs

Applications

11. Bias and Fairness ← *You are here!* ✓
12. Actual Causality
13. Alternative Perspectives ← [30. Jan] *Lecture survey*
14. Meta-Causality

15. Recap & Questions ← [13. Feb] *Exercise exam released*

Course Overview

Pearl. Caus.

1. Introduction
2. Probabilities and Bayesian Networks
3. Structural Causal Models
4. do-Calculus
5. Causal Discovery
6. Uncertainty

Repr. & ML

7. Causal Abstractions
8. Neuro-Causal Models
9. Causal Representation Learning
10. Causality and LLMs

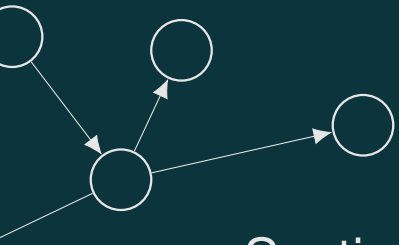
Applications

11. Bias and Fairness ← *You are here!* ✓
12. Actual Causality
13. Alternative Perspectives ← [30. Jan] *Lecture survey*
14. Meta-Causality

15. Recap & Questions ← [13. Feb] *Exercise exam released*

Exam Date

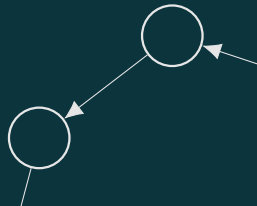
18. March
12:30-13:30



Section

1

Bias & Fairness



Associational Biases in LLM

Asking ChatGPT to complete the sentence:

1. The doctor told me that [MASK] would be on vacation next week.
2. The secretary told me that [MASK] would be on vacation next week.

McMilin, E., Selection Collider Bias in Large Language Models. In UAI 2022 Workshop on Causal Representation Learning.

Associational Biases in LLM

Asking ChatGPT to complete the sentence:

1. The doctor told me that HE(3/4)/THEY(1/4) would be on vacation next week.
2. The secretary told me that SHE(3/4)/THEY(1/4) would be on vacation next week.

McMilin, E., Selection Collider Bias in Large Language Models. In UAI 2022 Workshop on Causal Representation Learning.

Associational Biases in LLM

Asking ChatGPT to complete the sentence:

1. The doctor told me that would be on vacation next week.
2. The secretary told me that would be on vacation next week.

Models trained on biased data are likely to replicate those biases! ⚡

McMilin, E., Selection Collider Bias in Large Language Models. In UAI 2022 Workshop on Causal Representation Learning.

Real World Applications

ML and automated decision making is increasingly deployed in areas of

- Medicine
- Jurisdiction
- Employment
- Finance
- ...

Real World Applications

ML and automated decision making is increasingly deployed in areas of

- Medicine
- Jurisdiction
- Employment
- Finance
- ...

We want our predictions to be free of biases and prejudices for ethical reasons.

Real World Applications

ML and automated decision making is increasingly deployed in areas of

- Medicine
- Jurisdiction
- Employment
- Finance
- ...

We want our predictions to be free of biases and prejudices for ethical reasons.

$\rightarrow$ In certain areas this might be required by law! 

Biases in Model Training

Biases might get introduced at different stages of model training:

Biases in Model Training

Biases might get introduced at different stages of model training:

- Bias–variance tradeoff: The model architecture determines the type of relations that a model can learn.

Biases in Model Training

Biases might get introduced at different stages of model training:

- Bias–variance tradeoff: The model architecture determines the type of relations that a model can learn.
- Observational Bias: Statistical effects that skew the training data and, thus, model predictions.

Biases in Model Training

Biases might get introduced at different stages of model training:

- Bias–variance tradeoff: The model architecture determines the type of relations that a model can learn.
- Observational Bias: Statistical effects that skew the training data and, thus, model predictions.
- Cognitive Biases: Human biases (stereotyping, hindsight bias, . . .) might be adopted from training data.

Biases in Model Training

Biases might get introduced at different stages of model training:

- Bias–variance tradeoff: The model architecture determines the type of relations that a model can learn.
- Observational Bias: Statistical effects that skew the training data and, thus, model predictions.
- Cognitive Biases: Human biases (stereotyping, hindsight bias, . . .) might be adopted from training data.

Prevent our AI models to succumb to the same mistakes we do!

Bias in Image Generation

“Generate a photo of a firefighter”:

Stable diffusion only generated white, male-appearing persons as firefighters.

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S. and Kersting, K., 2025. Auditing and instructing text-to-image generation models on fairness. AI and Ethics, 5(3), pp.2103-2123.

Bias in Image Generation

“Generate a photo of a firefighter”:

Stable diffusion only generated white, male-appearing persons as firefighters.

→ Might reinforce biases! ⚡

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S. and Kersting, K., 2025. Auditing and instructing text-to-image generation models on fairness. AI and Ethics, 5(3), pp.2103-2123.

Bias in Image Generation - Countermeasures

“Generate a photo of a firefighter”:

Goal: introduce mechanisms to counteract biased generations.

Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S. and Kersting, K., 2025. Auditing and instructing text-to-image generation models on fairness. AI and Ethics, 5(3), pp.2103-2123.

Unfair Ground Truth

What if our ground truth data is already biased?

Historical bias: Historically, men had better access to education. If we were to take historic data as the ground truth for our predictions, our models will continue to reproduce these biases.

Unfair Ground Truth

What if our ground truth data is already biased?

Historical bias: Historically, men had better access to education. If we were to take historic data as the ground truth for our predictions, our models will continue to reproduce these biases.

Goal: introduce mechanisms to counteract biased generations.

Many more Biases…

Hindsight Bias Model Bias Association Bias
Participation Bias Selection Bias Sampling Bias Confirmation Bias
Measurement Bias Observation Bias Stereotyping
Survivorship Bias Gambler's fallacy Anchoring Bias
Gender Bias Outcome bias

Why Fairness?

Simply optimizing models for prediction accuracy might induce or carry over biases from existing data.

Why Fairness?

Simply optimizing models for prediction accuracy might induce or carry over biases from existing data.

Conflict of interest:

$\rightarrow$ “A bank could boost its profits by denying loans to poor people.” ⚡

$\rightarrow$ Boosting performance $\not\Leftrightarrow$ being fair

Why Fairness?

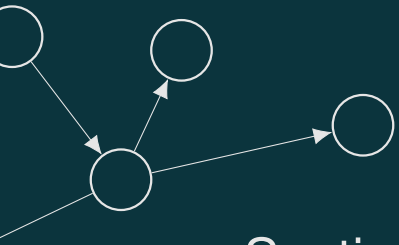
Simply optimizing models for prediction accuracy might induce or carry over biases from existing data.

Conflict of interest:

$\rightarrow$ "A bank could boost its profits by denying loans to poor people." ⚡

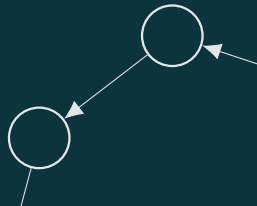
$\rightarrow$ Boosting performance $\not\leftrightarrow$ being fair

Build models that make accurate predictions while being fair!



Section 2

Bias in Causal Models



Graphical Modeling of Biases

Why to consider causal models for tackling bias?

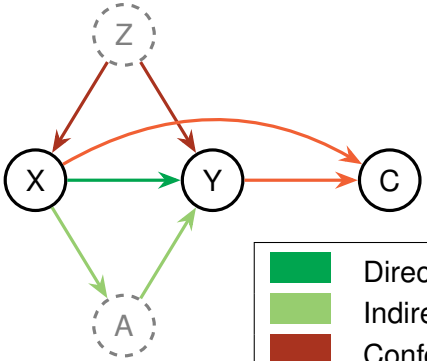
Causal models are “white-box”:

- Causal models tell us how effects propagate.
- Allow us to inspect and reason about the occurrence of biases.
- Estimate (and hopefully correct for) biasing factors!

Causal Effect Estimation

Remember causal effect estimation: identify the true (total) causal effect?

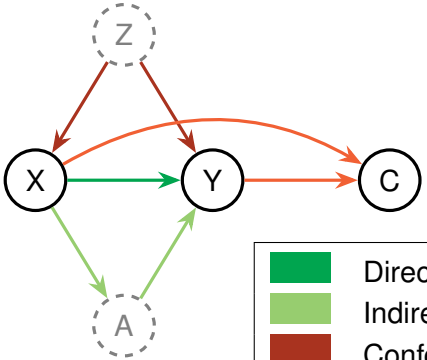
→ Correct for the influence of all non-causal paths in the estimation of $p(Y|do(X))$.



- Direct Causal Effect ($X \rightarrow Y$)
- Indirect Causal Effect ($X \rightarrow A \rightarrow Y$)
- Confounding Path ($X \leftarrow Z \rightarrow Y$; *Correlational!*)
- Collider Path ($X \rightarrow C \leftarrow Y$; *Correlational!*)

Causal Effect Estimation

Remember causal effect estimation: identify the true (total) causal effect?
→ Correct for the influence of all non-causal paths in the estimation of $p(Y|do(X))$.



Failing to adjust for any of the
non-causal paths will bias the estimate!

- Direct Causal Effect ($X \rightarrow Y$)
- Indirect Causal Effect ($X \rightarrow A \rightarrow Y$)
- Confounding Path ($X \leftarrow Z \rightarrow Y$; *Correlational!*)
- Collider Path ($X \rightarrow C \leftarrow Y$; *Correlational!*)

Collider Bias / Berkson's paradox

Scenario: Collecting data for a curated "Top Photos" gallery ($Z = 1$). Images are selected for the gallery if they are either *high-resolution* or show a *rare bird species*.

Common bird, high-resolution image.

Rare bird, low-resolution image.

Sparrow image: https://commons.wikimedia.org/wiki/File:House_Sparrow(Passer_domesticus).jpg

Collider Bias / Berkson's paradox

Scenario: Collecting data for a curated "Top Photos" gallery ($Z = 1$). Images are selected for the gallery if they are either *high-resolution* or show a *rare bird species*.

Training a model on this dataset may learn a *spurious correlation*. The data may imply that "rarity" implies "low resolution".

Common bird, high-resolution image.

Rare bird, low-resolution image.

Sparrow image: https://commons.wikimedia.org/wiki/File:House_Sparrow(Passer_domesticus).jpg

Collider Bias / Berkson's paradox

Scenario: Collecting data for a curated "Top Photos" gallery ($Z = 1$). Images are selected for the gallery if they are either *high-resolution* or show a *rare bird species*.

Training a model on this dataset may learn a *spurious correlation*. The data may imply that "rarity" implies "low resolution".

Deploying the trained model in a real-world setting will likely flag many low-resolution images as rare birds.

Common bird, high-resolution image.

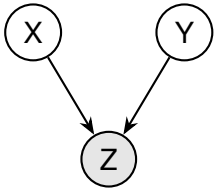
Rare bird, low-resolution image.

Sparrow image: https://commons.wikimedia.org/wiki/File:House_Sparrow(Passer_domesticus).jpg

Collider Bias / Berkson's paradox

A collider implies $P(X, Y) = P(X)P(Y)$,
but $P(X, Y|Z) \neq P(X|Z)P(Y|Z)$.

Explaining away: Within the scope of the dataset ($Z = 1$), observing $X = 1$ reduces the probability of observing $Y = 1$ at the same time (and vice versa).

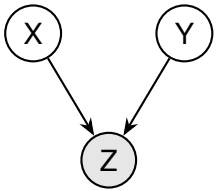


$Z := X \vee Y$

Collider Bias / Berkson's paradox

A collider implies $P(X, Y) = P(X)P(Y)$,
but $P(X, Y|Z) \neq P(X|Z)P(Y|Z)$.

Explaining away: Within the scope of the dataset ($Z = 1$), observing $X = 1$ reduces the probability of observing $Y = 1$ at the same time (and vice versa).



$Z := X \vee Y$

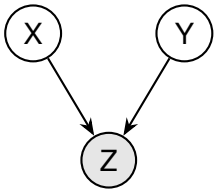
$$P(Y = 1|Z = 1, X = 1) = \frac{P(Z = 1|X = 1, Y = 1)P(Y = 1|X = 1)}{P(Z = 1|X = 1)} = P(Y = 1)$$

$$P(Y = 1|Z = 1, X = 0) = \frac{P(Z = 1|X = 0, Y = 1)P(Y = 1|X = 0)}{P(Z = 1|X = 0)} = 1$$

$$\Rightarrow \quad P(Y = 1|Z = 1, X = 1) = P(Y = 1) \leq 1 = P(Y = 1|Z = 1, X = 0)$$

Collider Bias / Berkson's paradox

A collider implies $P(X, Y) = P(X)P(Y)$,
but $P(X, Y|Z) \neq P(X|Z)P(Y|Z)$.



$Z := X \vee Y$

Explaining away: Within the scope of the dataset ($Z = 1$), observing $X = 1$ reduces the probability of observing $Y = 1$ at the same time (and vice versa).

$$P(Y = 1|Z = 1, X = 1) = \frac{P(Z = 1|X = 1, Y = 1)P(Y = 1|X = 1)}{P(Z = 1|X = 1)} = P(Y = 1)$$

$$P(Y = 1|Z = 1, X = 0) = \frac{P(Z = 1|X = 0, Y = 1)P(Y = 1|X = 0)}{P(Z = 1|X = 0)} = 1$$

$$\Rightarrow \quad P(Y = 1|Z = 1, X = 1) = P(Y = 1) \leq 1 = P(Y = 1|Z = 1, X = 0)$$

Takeaway: *Selecting data based on 'quality' or 'interest' can make previously independent features appear dependent.*

Simpson Paradox

Simply regressing the effect of *hours of exercise* on *cholesterol levels* over the whole population predicts a negative effect of exercise on cholesterol levels…

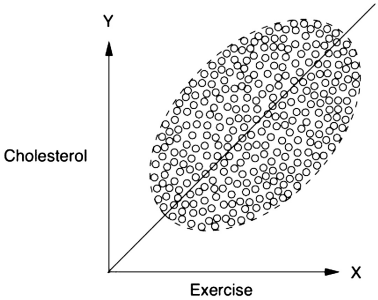
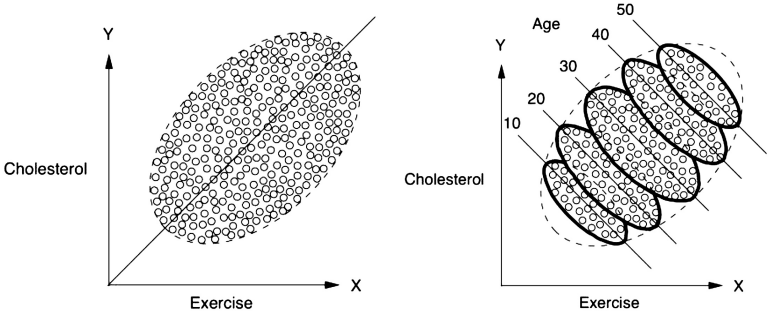


Figure: Pearl, J., Mackenzie D., 2018. The book of why: the new science of cause and effect, Basic Books.

Simpson Paradox

Simply regressing the effect of *hours of exercise* on *cholesterol levels* over the whole population predicts a negative effect of exercise on cholesterol levels…



… however, adjusting by *age group* reveals the true *positive* causal effect.

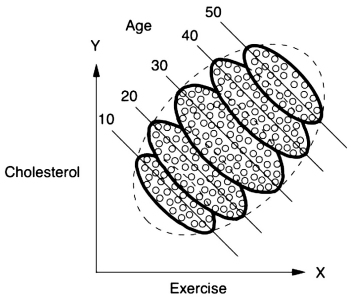
Figure: Pearl, J., Mackenzie D., 2018. The book of why: the new science of cause and effect, Basic Books.

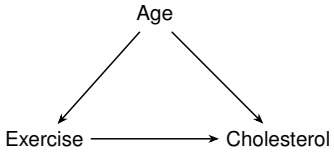
Simpson Paradox

Observation: Exercise has a positive effect at every age group, but is harmful on over the whole population!?

Hypothetical scenario:

- ‘Exercising influences cholesterol levels.’
- ‘Age influences cholesterol levels.’
- ‘Older people exercise more.’





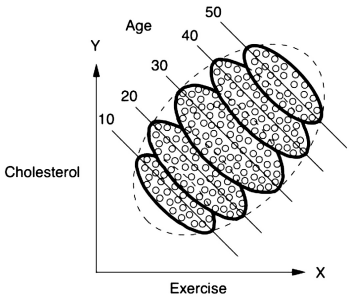
Simpson Paradox

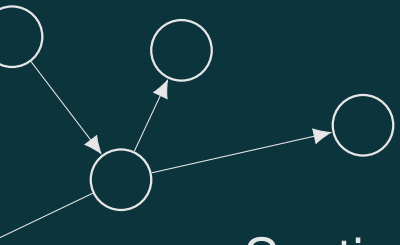
Observation: Exercise has a positive effect at every age group, but is harmful on over the whole population!?

Hypothetical scenario:

- 'Exercising influences cholesterol levels.' (direct causal effect)
- 'Age influences cholesterol levels.' (opens backdoor path)
- 'Older people exercise more.' (opens backdoor path)

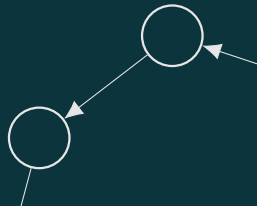
Typical example of backdoor adjustment.
$\rightarrow$ do-calculus tells us to correct for age.





Section 3

Fairness



Modeling Decision Processes

A decision process for a particular task T is defined over the following spaces:

- **Construct space** $\mathcal{CS} = (P, d_P)$: space over individuals $p \in P$.
- **Observed space** $\mathcal{OS} = (\hat{P}, \hat{d})$:
  Actual recorded/observed data of individuals, $\hat{p} := g(p)$.
- **Decision Space** $\mathcal{DS} = (O, d_O)$:
  Space representing the actual outcomes, $o := t(p)$.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. "Communications of the ACM 64.4 (2021): 136-143.

Modeling Decision Processes

A decision process for a particular task T is defined over the following spaces:

- **Construct space** $\mathcal{CS} = (P, d_P)$: space over individuals $p \in P$.
- **Observed space** $\mathcal{OS} = (\hat{P}, \hat{d})$:
  Actual recorded/observed data of individuals, $\hat{p} := g(p)$.
- **Decision Space** $\mathcal{DS} = (O, d_O)$:
  Space representing the actual outcomes, $o := t(p)$.

Decision space	*Construct space*	*Observed space*
Performance in college	Success in High School	GPA
Employee Productivity	Knowledge of job	Number of Years of Experience

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. "Communications of the ACM 64.4 (2021): 136-143.

Modeling Decision Processes

A decision process for a particular task T is defined over the following spaces:

- **Construct space** $\mathcal{CS} = (P, d_P)$: space over individuals $p \in P$.
- **Observed space** $\mathcal{OS} = (\hat{P}, \hat{d})$:
  Actual recorded/observed data of individuals, $\hat{p} := g(p)$.
- **Decision Space** $\mathcal{DS} = (O, d_O)$:
  Space representing the actual outcomes, $o := t(p)$.

Decision space	*Construct space*	*Observed space*
Performance in college	Success in High School	GPA
Employee Productivity	Knowledge of job	Number of Years of Experience

Challenge: Properties of the construct space are commonly not directly observed.
$\rightarrow$ Properties of the observed space are used as proxies instead.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. "Communications of the ACM 64.4 (2021): 136-143.

Fairness Metrics

Fairness is measured as the divergence in distance between individuals in construct space and the actual decision space.

(ϵ, ϵ')-Fairness [Frieder et al., 2021]

A mapping $f : \mathcal{CS} \to \mathcal{DS}$ is said to be fair if objects that are close in $\mathcal{CS}$ are also close in $\mathcal{DS}$. Specifically, fix two thesholds ϵ, ϵ'. Then f is defined as (ϵ, ϵ')-fair if for any $x, y \in P$,

$$d_P(x, y) \leq \epsilon \Rightarrow d_O(f(x), f(y)) \leq \epsilon'$$

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making." Communications of the ACM 64.4 (2021): 136-143.

Fairness Metrics

Fairness is measured as the divergence in distance between individuals in construct space and the actual decision space.

(ϵ, ϵ')-Fairness [Frieder et al., 2021]

A mapping $f : \mathcal{CS} \to \mathcal{DS}$ is said to be fair if objects that are close in $\mathcal{CS}$ are also close in $\mathcal{DS}$. Specifically, fix two thesholds ϵ, ϵ'. Then f is defined as (ϵ, ϵ')-fair if for any $x, y \in P$,

$$d_P(x, y) \leq \epsilon \Rightarrow d_O(f(x), f(y)) \leq \epsilon'$$

Note: The definition restricts the permitted divergence between construct and decision space, but defers the problem of fairness to defining a 'fair' construct space.

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making." Communications of the ACM 64.4 (2021): 136-143.

Protected Attributes

Some attributes might be *protected*.

$\rightarrow$ Protected attributes ($\mathbf{A} \subset \mathbf{X}$) must not be used for making predictions!

Protected attributes might include:

1. Age
2. Cultural/ethnic background
3. Sexual orientation
4. Religion or belief
5. ...

They are usually defined by ethical or legal considerations.

Protected Attributes

Some attributes might be *protected*.

$\rightarrow$ Protected attributes ($\mathbf{A} \subset \mathbf{X}$) must not be used for making predictions!

Protected attributes might include:

1. Age
2. Cultural/ethnic background
3. Sexual orientation
4. Religion or belief
5. ...

They are usually defined by ethical or legal considerations.

We would like our predictions $\hat{Y}$ of some target label $Y \in \mathbf{X}$ to be invariant to $\mathbf{A}$.

Demographic Parity

Simple idea:

An outcome should be equally likely, independent of the protected attributes.

Demographic Parity

$$\forall a \in \mathcal{A}.\ P(\hat{Y}|A = a) = P(\hat{Y}|A = a')$$

Demographic Parity

Simple idea:

An outcome should be equally likely, independent of the protected attributes.

Demographic Parity

$$\forall a \in \mathcal{A}.\ P(\hat{Y}|A = a) = P(\hat{Y}|A = a')$$

Example:

"Employ people independent of their parental status."
"Grant loans to all ethnic groups equally."
"Hire people independent of their religion."
…

Demographic Parity

Simple idea:

An outcome should be equally likely, independent of the protected attributes.

Measuring the disparity between terms quantifies the (un)fairness of the decision.

Demographic Parity

$$\forall a \in \mathcal{A}.\ P(\hat{Y}|A = a) = P(\hat{Y}|A = a')$$

Example:

"Employ people independent of their parental status."
"Grant loans to all ethnic groups equally."
"Hire people independent of their religion."
…

Unfair Demographic Parity

Demographic parity does not prevent us from being unfair:

- Grant college admission to 50% of the most qualified students of group A.
- Grant college admission to 50% of randomly selected students of group B.

$\rightarrow$ Group B is likely to be less successful.

Unfair Demographic Parity

Demographic parity does not prevent us from being unfair:

- Grant college admission to 50% of the most qualified students of group A.
- Grant college admission to 50% of randomly selected students of group B.

$\rightarrow$ Group B is likely to be less successful.

Insight: Demographic parity ($P(\hat{Y}|A = a) = P(\hat{Y}|A = a')$) only cares about the assigned prediction, not whether the label is correct or not.

Unfair Demographic Parity

Demographic parity does not prevent us from being unfair:

- Grant college admission to 50% of the most qualified students of group A.
- Grant college admission to 50% of randomly selected students of group B.

$\rightarrow$ Group B is likely to be less successful.

Insight: Demographic parity ($P(\hat{Y}|A = a) = P(\hat{Y}|A = a')$) only cares about the assigned prediction, not whether the label is correct or not.

Idea: Involve the true label Y for measuring prediction disparities among groups.

Equality of Opportunity

We would like to have balance the percentage of positive outcomes (same *True Positive Rate*) for all $a \in \mathcal{A}$.

Equality of opportunity

$$\forall a \in \mathcal{A}.\ P(\hat{Y} | Y = 1, A = a) = P(\hat{Y} | Y = 1, A = a')$$

In plain words: "If two students are both qualified for a course ($Y = 1$), do they have the same chance of being approved ($\hat{Y} = 1$) regardless of their origin or gender?"

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

Unfair Equality of Opportunity

Equality of Opportunity does not prevent us from being unfair!

General Bias: A *False Positive* means an innocent person stays in jail longer. If one group has a higher False Positive Rate than the others, the model is systematically "over-punishing" that group. ⚡

Unfair Equality of Opportunity

Equality of Opportunity does not prevent us from being unfair!

General Bias: A *False Positive* means an innocent person stays in jail longer. If one group has a higher False Positive Rate than the others, the model is systematically "over-punishing" that group. ϟ

Under Constraints: Universities have a limited number of seats. If the model "mistakenly" admits students of group A more often, it might be taking away seats from qualified students of other groups, reducing their chances of admission. ϟ

Equalized odds

Idea: Enforce an equal true-positive *and false-positive rate* for all values of A.

Equalized odds

$$\forall a \in A, y \in \{0, 1\}.\ P(\hat{Y} | Y = y, A = a) = P(\hat{Y} | Y = y, A = a')$$

Problems with Correlational Fairness

So far, every notion of fairness balanced some probability between subgroups.

Problems with Correlational Fairness

So far, every notion of fairness balanced some probability between subgroups.

$\rightarrow$ These metrics are nice from a ‘global’ point of view, but don’t help from an individual perspective.

Problems with Correlational Fairness

So far, every notion of fairness balanced some probability between subgroups.

$\rightarrow$ These metrics are nice from a 'global' point of view, but don't help from an individual perspective.

Upon receiving his rejection letter from university Tom asks himself:
"Would I have been accepted if I had been a different age?"

Problems with Correlational Fairness

So far, every notion of fairness balanced some probability between subgroups.

→ These metrics are nice from a 'global' point of view, but don't help from an individual perspective.

Upon receiving his rejection letter from university Tom asks himself:
"Would I have been accepted if I had been a different age?"

This is a classical counterfactual question!

Classes of Fairness Metrics

Previous definitions of fairness considered groups
$\rightarrow$ This potentially opens up ways to discriminate against individuals.

Group Fairness:
Requires individuals from different subgroups to be treated equally (according to some metric and set of protected attributes).

Individual Fairness:
Requires similar individuals to be treated equally, (regardless of their group membership).

Counterfactual Fairness

Idea: Require the same potential outcome for individuals, independent of $\mathbf{A}$.

Counterfactual Fairness

$$P(\hat{Y}_{\mathbf{A}=\mathbf{a}}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a}) = P(\hat{Y}_{\mathbf{A}=\mathbf{a}'}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a})$$

Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Counterfactual Fairness

Idea: Require the same potential outcome for individuals, independent of **A**.

Counterfactual Fairness

$$P(\hat{Y}_{\mathbf{A}=\mathbf{a}}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a}) = P(\hat{Y}_{\mathbf{A}=\mathbf{a}'}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a})$$

Example: People with the same qualifications should be given the same opportunity for college admission, independent of their protected attributes.

Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Counterfactual Fairness

Idea: Require the same potential outcome for individuals, independent of $\mathbf{A}$.

Counterfactual Fairness

$$P(\hat{Y}_{\mathbf{A}=\mathbf{a}}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a}) = P(\hat{Y}_{\mathbf{A}=\mathbf{a}'}(\mathbf{U}) | \mathbf{V}=\mathbf{v}, \mathbf{A}=\mathbf{a})$$

Example: People with the same qualifications should be given the same opportunity for college admission, independent of their protected attributes.

Counterfactual Fairness:

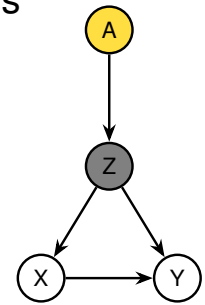
→ compares outcomes *per individual*.

→ does not depend on some (possibly biased) ground thruth label Y.

Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Counterfactual Fairness in Graphical Models and NNs

Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

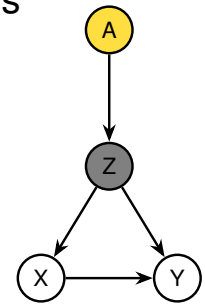


Counterfactual Fairness in Graphical Models and NNs

Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

$\rightarrow$ This, however, requires graph specific estimands, e.g.:

$$P(Y|do(X=x)) = \sum_{z \in \mathcal{Z}} P(Y|X=x, Z=z)P(Z=z)$$



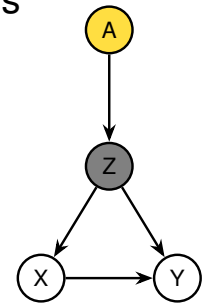
Counterfactual Fairness in Graphical Models and NNs

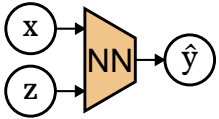
Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

$\rightarrow$ This, however, requires graph specific estimands, e.g.:

$$P(Y|do(X=x)) = \sum_{z \in \mathcal{Z}} P(Y|X=x, Z=z)P(Z=z)$$

Neural models commonly correspond to a single, joint regression function: $\hat{Y} := f(X, Z)$.





Counterfactual Fairness in Graphical Models and NNs

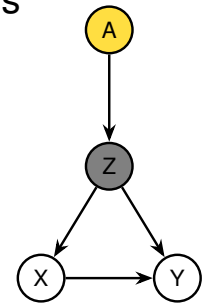
Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

$\rightarrow$ This, however, requires graph specific estimands, e.g.:

$$P(Y|do(X=x)) = \sum_{z \in \mathcal{Z}} P(Y|X=x, Z=z)P(Z=z)$$

Neural models commonly correspond to a single, joint regression function: $\hat{Y} := f(X, Z)$.

$\rightarrow$ Is oblivious to the causal graphical structure.

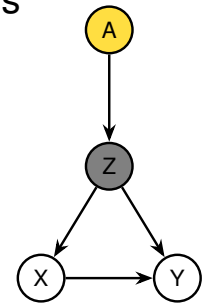


Counterfactual Fairness in Graphical Models and NNs

Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

$\rightarrow$ This, however, requires graph specific estimands, e.g.:

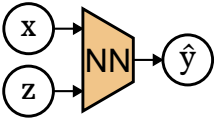
$$P(Y|do(X=x)) = \sum_{z \in \mathcal{Z}} P(Y|X=x, Z=z)P(Z=z)$$



Neural models commonly correspond to a single, joint regression function: $\hat{Y} := f(X, Z)$.

$\rightarrow$ Is oblivious to the causal graphical structure.

$\rightarrow$ Does not differentiate between causes X and the conditioning set Z.

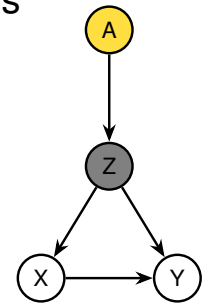


Counterfactual Fairness in Graphical Models and NNs

Given causal structural knowledge, one can correct for the effects of the protected attributes via $P(\hat{Y}|do(X))$ and $X \perp\!\!\!\perp \mathbf{A}$.

→ This, however, requires graph specific estimands, e.g.:

$$P(Y|do(X=x)) = \sum_{z \in \mathcal{Z}} P(Y|X=x, Z=z)P(Z=z)$$

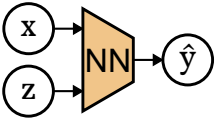


Neural models commonly correspond to a single, joint regression function: $\hat{Y} := f(X, Z)$.

→ Is oblivious to the causal graphical structure.

→ Does not differentiate between causes X and the conditioning set Z.

→ NN do not approximate the true causal estimand. ⚡



Counterfactual Fairness for ML Training

Lemma 1 [Kusner et al, 2017].

Let $\mathcal{G}$ be the causal graph of a given model $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$. Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A.

$\rightarrow$ Provides a condition which attributes can be safely used for model training.

Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Counterfactual Fairness for ML Training

Lemma 1 [Kusner et al, 2017].

Let $\mathcal{G}$ be the causal graph of a given model $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathcal{P}_{\mathbf{U}})$. Then $\hat{Y}$ will be counterfactually fair if it is a function of the non-descendants of A.

$\rightarrow$ Provides a condition which attributes can be safely used for model training.

Proof (informal): For any non-descendants $\mathbf{W}$ we infer the same counterfactual values $\mathbf{W}^*_{A \leftarrow a}$ since they are independent from $\mathbf{A}$. Hence, using $\mathbf{W}$ for training our classifier $\hat{Y}$ is invariant to the counterfactual values of $\mathbf{A}$.

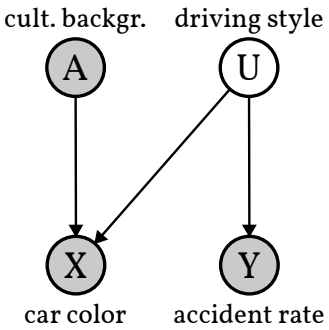
Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Counterfactual Fairness Example

Assume the following scenario:

Some culture ($A = a^*$) prefers red cars ($X =$ 'red').
$\rightarrow$ while having an *equal* accident rate $Y = y^0$.

Drivers with an aggressive driving style ($U =$ '++') prefer red cars.
$\rightarrow$ while having an *increased* accident rate $Y = y^{\uparrow}$.



Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

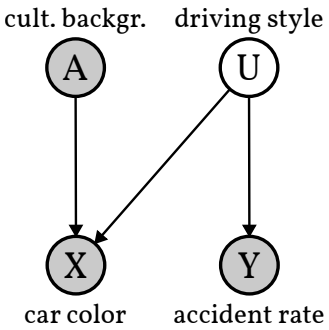
Counterfactual Fairness Example

Assume the following scenario:

Some culture ($A = a^*$) prefers red cars ($X =$ 'red').
$\rightarrow$ while having an *equal* accident rate $Y = y^0$.

Drivers with an aggressive driving style ($U =$ '++')
prefer red cars.
$\rightarrow$ while having an *increased* accident rate $Y = y^{\uparrow}$.

Regressing on X is unfair towards culture $A = a^*$,
since Y and A are no longer d-separated.



Kusner, M.J., Loftus, J., Russell, C. and Silva, R., 2017. Counterfactual fairness. Advances in neural information processing systems, 30.

Fairness through Unawareness

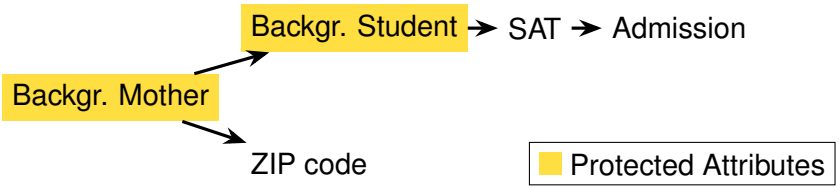
Counterfactual fairness is sometimes referred to as ‘fairness through unawareness’.

$\rightarrow$ This should not be mistaken as to ‘not care’ about protected attributes.

Fairness through Unawareness

Counterfactual fairness is sometimes referred to as ‘fairness through unawareness’.
→ This should not be mistaken as to ‘not care’ about protected attributes.

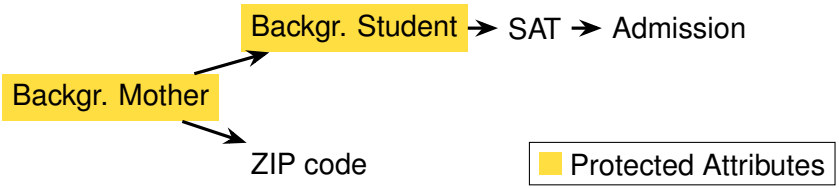
Unprotected attributes, like the ZIP code or name, might be highly correlated and induce bias again:



Fairness through Unawareness

Counterfactual fairness is sometimes referred to as ‘fairness through unawareness’.
→ This should not be mistaken as to ‘not care’ about protected attributes.

Unprotected attributes, like the ZIP code or name, might be highly correlated and induce bias again:



→ One needs to explicitly *acquire knowledge on the causal structure* to guarantee the independence between protected attributes and the final decision: $\hat{Y} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{A}$.

Incompatibility of Fairness Metrics

Many fairness metrics are mutually incompatible.
$\rightarrow$ They can not be satisfied at the same time.

https://developers.google.com/machine-learning/glossary/responsible-ai#incompatibility_of_fairness_metrics

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making." Communications of the ACM 64.4 (2021): 136-143.

Incompatibility of Fairness Metrics

Many fairness metrics are mutually incompatible.
$\rightarrow$ They can not be satisfied at the same time.

The exists no single universal metric for quantifying fairness.
$\rightarrow$ Individual- versus population-level fairness concerns.
$\rightarrow$ Predictability versus fairness tradeoffs.

https://developers.google.com/machine-learning/glossary/responsible-ai#incompatibility_of_fairness_metrics

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making." Communications of the ACM 64.4 (2021): 136-143.

Incompatibility of Fairness Metrics

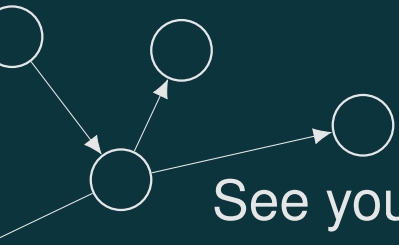
Many fairness metrics are mutually incompatible.
$\rightarrow$ They can not be satisfied at the same time.

The exists no single universal metric for quantifying fairness.
$\rightarrow$ Individual- versus population-level fairness concerns.
$\rightarrow$ Predictability versus fairness tradeoffs.

The selection of fairness metrics should be a deliberate choice and decided situationally depending on the goals and context.

https://developers.google.com/machine-learning/glossary/responsible-ai#incompatibility_of_fairness_metrics

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. "The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making." Communications of the ACM 64.4 (2021): 136-143.



See you next week!