Generating Counterfactuals for Fairness using VAE

Ayan Majumdar, Preethi Lahoti, Junaid Ali, Till Speicher,
Isabel Valera, Krishna Gummadi

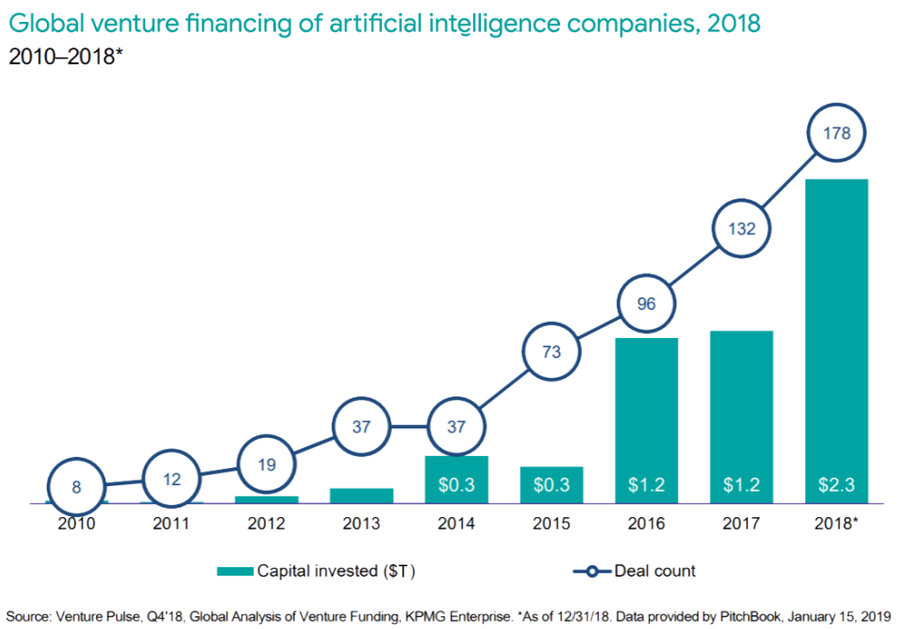
MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

max planck institut
informatik

UNIVERSITÄT
DES
SAARLANDES

ML and our society

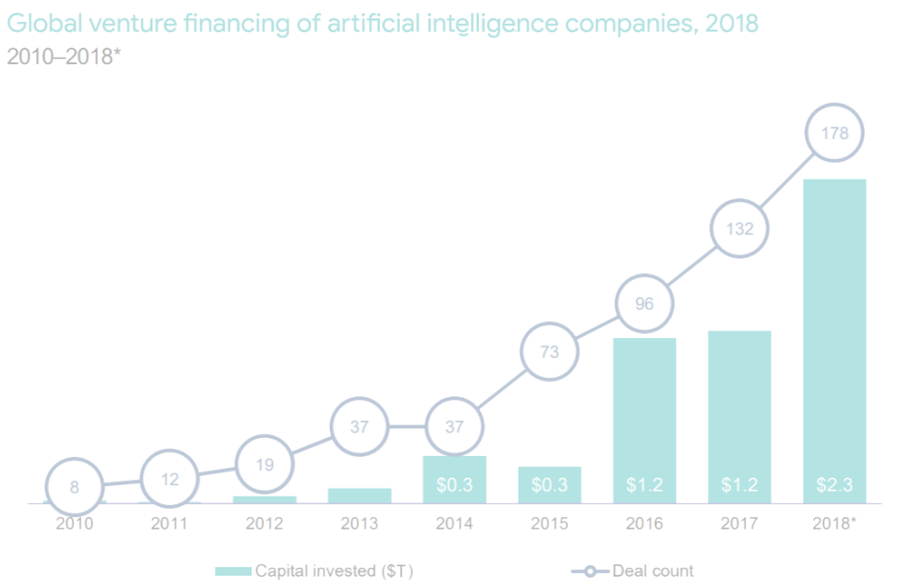
Data-driven ML algorithms **heavily** deployed in today's tech industry.



Increased industry financing for AI and ML

ML and our society

Data-driven ML algorithms more **pervasive** in today's society.



Increased industry financing for AI and ML

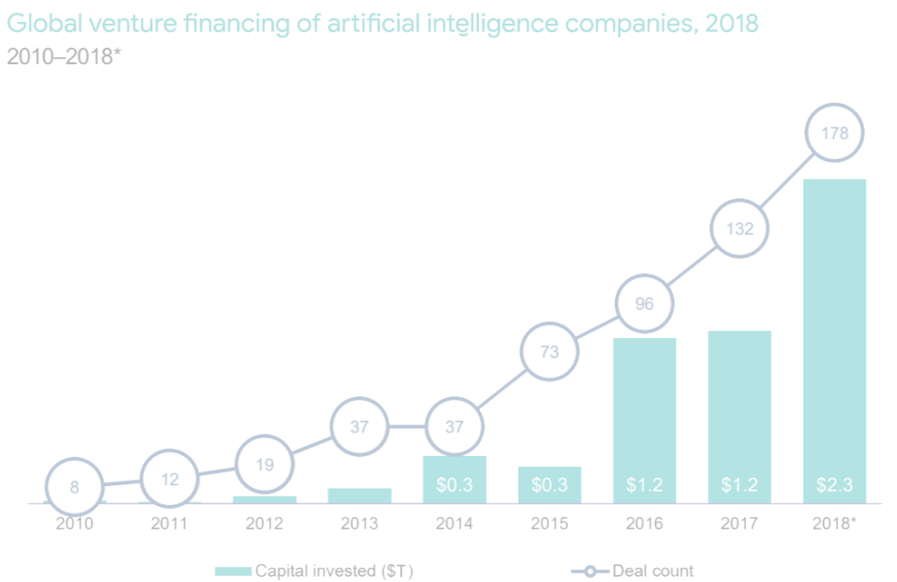
ML and our society

Data-driven ML algorithms more **pervasive** in today's society.

NORTHPOINTE

PredPol
Predict Crime in Real Time.
PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

ShotSpotter®

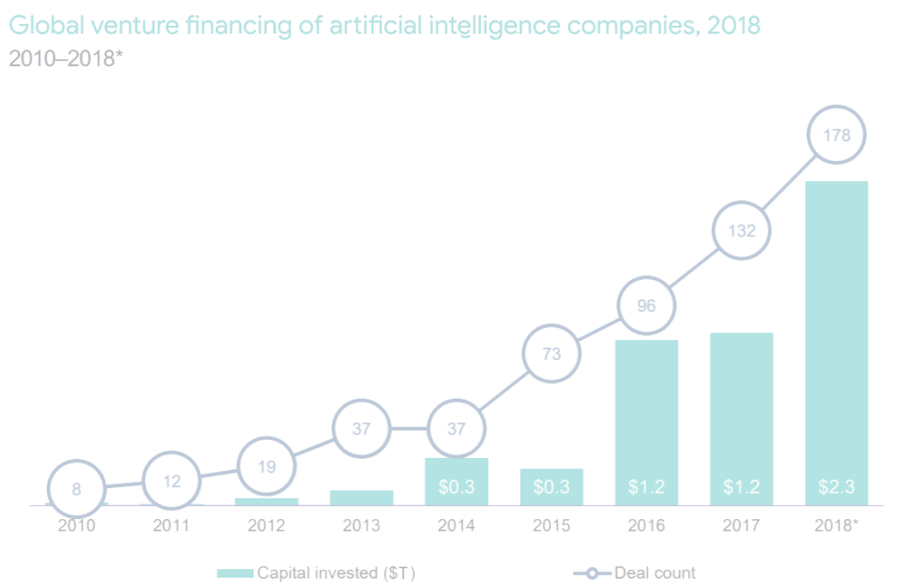


Source: Venture Pulse, Q4'18, Global Analysis of Venture Funding, KPMG Enterprise. *As of 12/31/18. Data provided by PitchBook, January 15, 2019

Increased industry financing for AI and ML

ML and our society

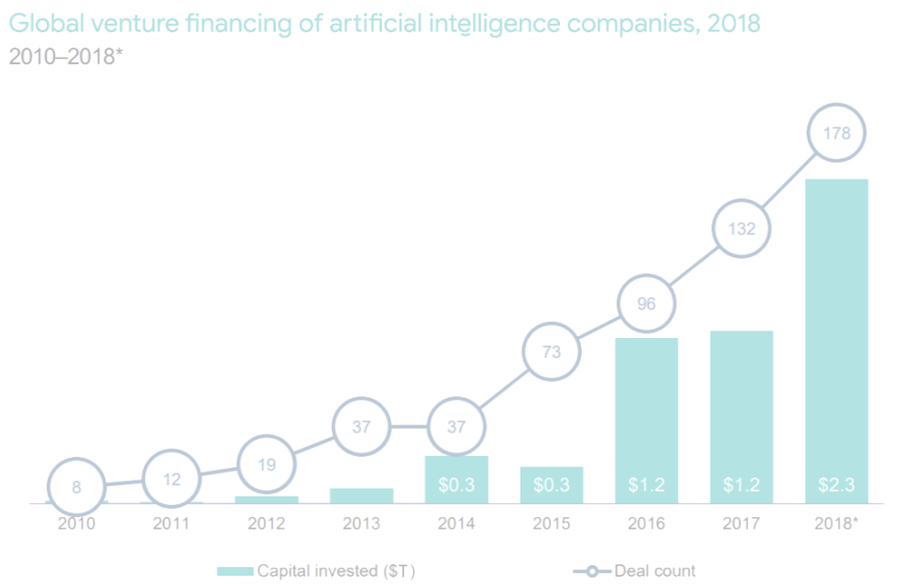
Data-driven ML algorithms more **pervasive** in today's society.



Increased industry financing for AI and ML

ML and our society

Data-driven ML algorithms more **pervasive** in today's society.



Increased industry financing for AI and ML

Fairness in ML systems

Studies have shown potential **bias** in deployed systems!

Business

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the higher credit score

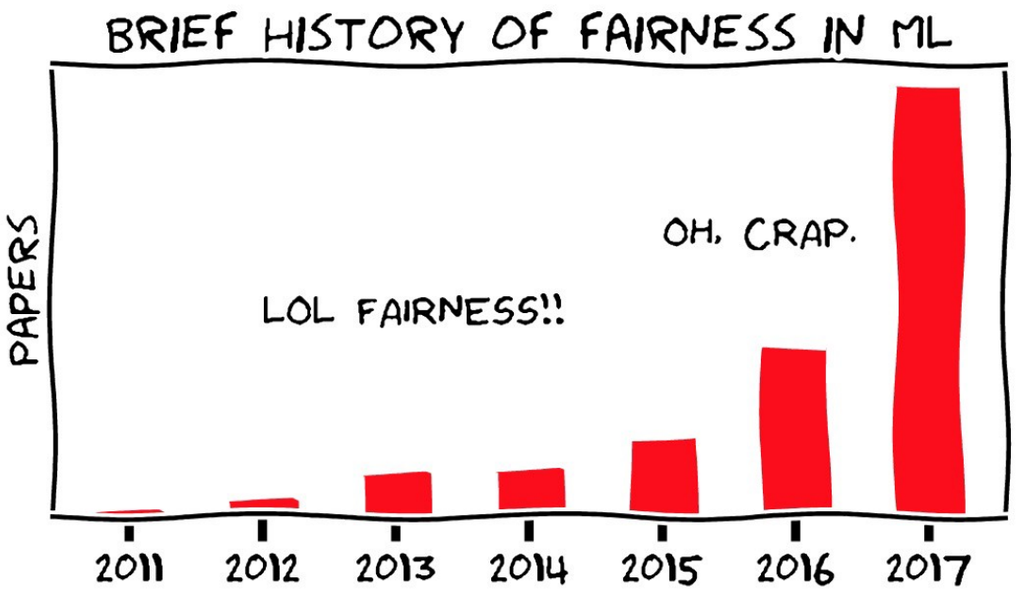
Why Amazon's Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program
OCTOBER 12, 2018 | 1:00 PM

TAGS: Women's Rights in the Workplace, Women's Rights, Privacy & Technology

Fairness in ML systems

Led to **extensive** research in the domain…



Credits: Moritz Hardt, CS 294-Fairness in Machine Learning

Business

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the higher credit score

Why Amazon's Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program
OCTOBER 12, 2018 | 1:00 PM

TAGS: Women's Rights in the Workplace, Women's Rights, Privacy & Technology

Existing notions of fairness

Many definitions

- Group: *Fairness through unawareness, demographic parity, etc.*
- Individual: *individual fairness*

Existing notions of fairness

Many definitions

- Group: *Fairness through unawareness, demographic parity, etc.*
- Individual: *individual fairness*

However…

Existing notions of fairness

Many definitions

— Group: *Fairness through unawareness, demographic parity, etc.*

— Individual: *individual fairness*

However…

- How sensitive attributes **cause** bias?
- How to **eliminate** bias?
- How to define **similar** individuals (*individual fairness*)?

Existing notions of fairness

Many definitions

— Group: *Fairness through unawareness, demographic parity, etc.*

— Individual: *individual fairness*

However…

- How sensitive attributes **cause** bias?
- How to **eliminate** bias?
- How to define **similar** individuals (*individual fairness*)?

Not clear!

💡Use causation in fairness!

💡Use causation in fairness!

Is the law school admission process fair?

💡Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3.

Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

- Had Jacob been white instead, would he had been accepted?

Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

- Had Jacob been white instead, would he had been accepted?
  — *counterfactual*

Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

- Had Jacob been white instead, would he had been accepted?
  — *counterfactual*
- Did Jacob's race **cause** him to get negative outcome?

Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

- Had Jacob been white instead, would he had been accepted?
  — *counterfactual*
- Did Jacob's race **cause** him to get negative outcome?
  — *counterfactual fairness* (*Kusner et al. 2017*)[1]

[1]*Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Use causation in fairness!

Is the law school admission process fair?

Jacob is a black male law school applicant. He scored 55 in LSAT and had UGPA 3.3. He was rejected.

- Had Jacob been white instead, would he had been accepted?
  — *counterfactual*
- Did Jacob's race **cause** him to get negative outcome?
  — *counterfactual fairness* (*Kusner et al. 2017*)[1]

Such questions of fairness need **counterfactual data**!

[1]*Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Counterfactuals, simply[1]

- **Change** Jacob's race
- Keep everything else **same**
- **Evaluate** decision outcome

[1] *Greiner et.al., "Causal effects of perceived immutable characteristics", The Review of Economics and Statistics, 2011*

Counterfactuals, ~~simply~~[1]

not so simple!

Simple approach[1]:

- **Change** Jacob's race
- Keep everything else **same**
- **Evaluate** decision outcome

→

Not enough!

- Very **limited** analysis
- No account for historical **bias** → data bias
  - *E.g. redlining affecting education*[2]

[1] *Greiner et.al., "Causal effects of perceived immutable characteristics", The Review of Economics and Statistics, 2011*
[2] *Michael Holzmann, "A Rotting Apple: Education Redlining in New York City", Schott Foundation for Public Education*

Counterfactuals: not so simple!

Simple approach[1]:

- **Change** Jacob's race
- Keep everything else **same**
- **Evaluate** decision outcome

→

Not enough!

- Very **limited** analysis
- No account for historical **bias** → data bias
  - *E.g. redlining affecting education*[2]

How do we estimate **bias caused** by **race**?

[1] *Greiner et.al., "Causal effects of perceived immutable characteristics", The Review of Economics and Statistics, 2011*
[2] *Michael Holzmann, "A Rotting Apple: Education Redlining in New York City", Schott Foundation for Public Education*

Counterfactuals: not so simple!

Simple approach[1]:

- **Change** Jacob's race
- Keep everything else **same**
- **Evaluate** decision outcome

→

Not enough!

- Very **limited** analysis
- No account for historical **bias**
  → data bias
  - *E.g. redlining affecting education*[2]

How do we estimate **bias caused** by **race**?

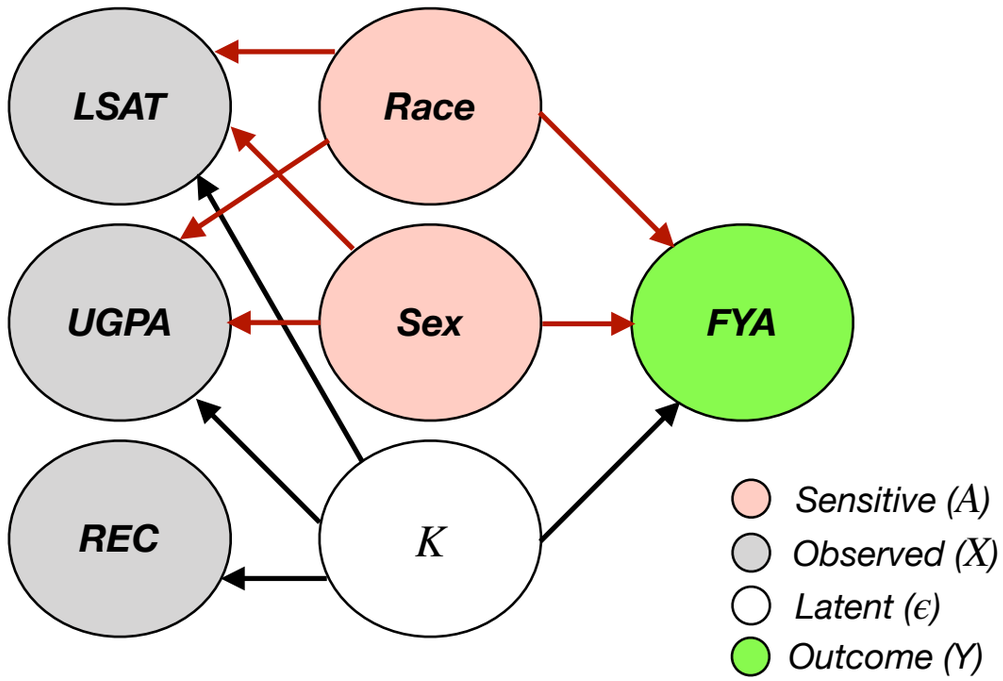
Need data **generative** process

[1]*Greiner et.al., "Causal effects of perceived immutable characteristics", The Review of Economics and Statistics, 2011*
[2]*Michael Holzmann, "A Rotting Apple: Education Redlining in New York City", Schott Foundation for Public Education*

Causal models

Causal models

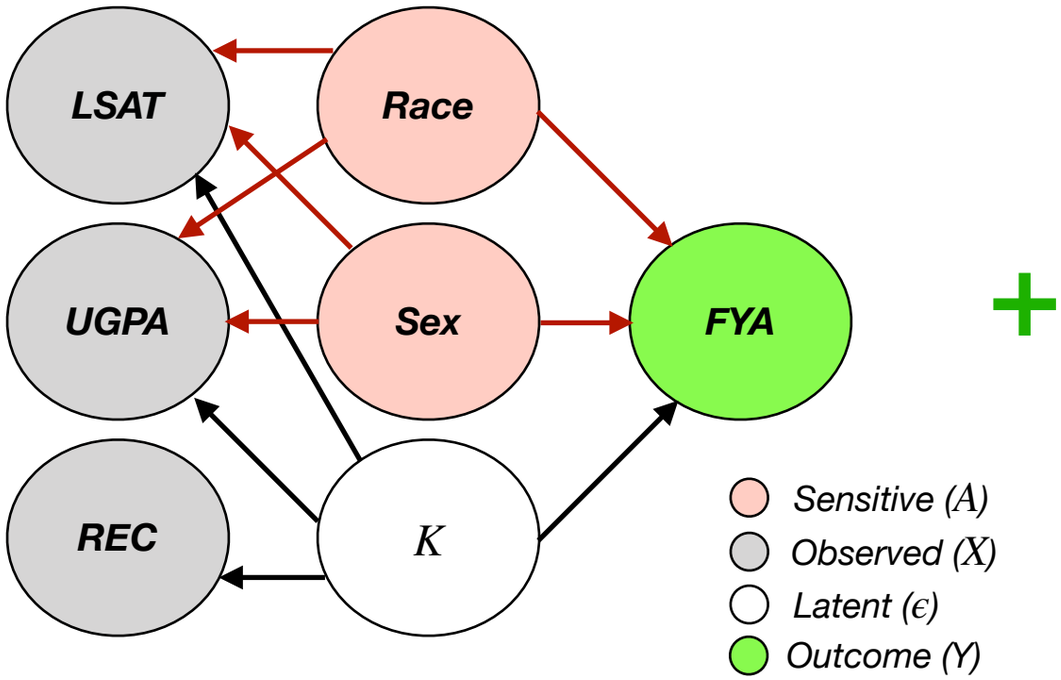
Causal graph



Relations between the features

Causal models

Causal graph



+

Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

$$\textbf{UGPA} := \mathcal{N}(b_G + w_G^R R + w_G^S S + w_G^K K, \sigma_G)$$

$$\textbf{FYA} := \mathcal{N}(w_F^R R + w_F^S S + w_F^K K, 1)$$

$$\textbf{REC} := \mathcal{N}(w_R^K K, 1)$$

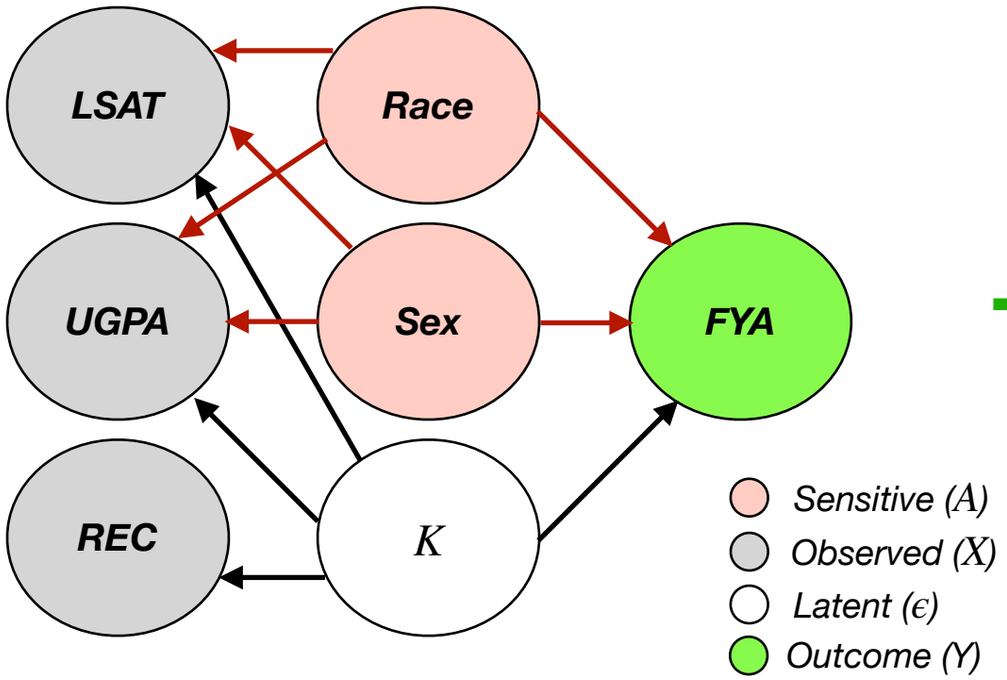
$$K \sim \mathcal{N}(0,1)$$

Relations between the features

Quantification of the relations

Causal models

Causal graph



+

Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

$$\textbf{UGPA} := \mathcal{N}(b_G + w_G^R R + w_G^S S + w_G^K K, \sigma_G)$$

$$\textbf{FYA} := \mathcal{N}(w_F^R R + w_F^S S + w_F^K K, 1)$$

$$\textbf{REC} := \mathcal{N}(w_R^K K, 1)$$

$$K \sim \mathcal{N}(0,1)$$

Relations between the features

Quantification of the relations

Need **complete** access! **Infeasible** in real settings.

Research Question

Can we generate **counterfactual** data for **fairness** in the absence of the whole **causal** model?

Research Question

Can we generate **counterfactual** data for **fairness** in the absence of the whole **causal** model?

1. Can we use the generated **counterfactuals** to audit existing predictive models?

Research Question

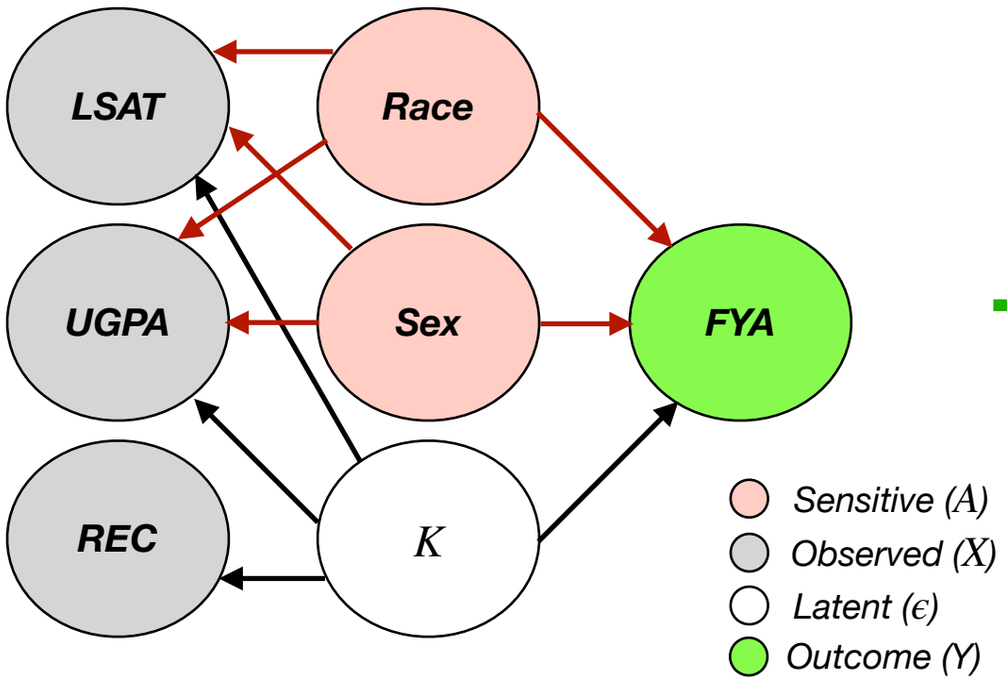
Can we generate **counterfactual** data for **fairness** in the absence of the whole **causal** model?

1. Can we use the generated **counterfactuals** to audit existing predictive models?

2. Can we use our approach to build a **fair** predictive model?

Background: Causal counterfactuals

Causal graph



+

Structural equations

$$\textbf{LSAT} := \mathcal{N}(\exp(b_L + w_L^R R + w_L^S S + w_L^K K), \sigma_L)$$

$$\textbf{UGPA} := \mathcal{N}(b_G + w_G^R R + w_G^S S + w_G^K K, \sigma_G)$$

$$\textbf{FYA} := \mathcal{N}(w_F^R R + w_F^S S + w_F^K K, 1)$$

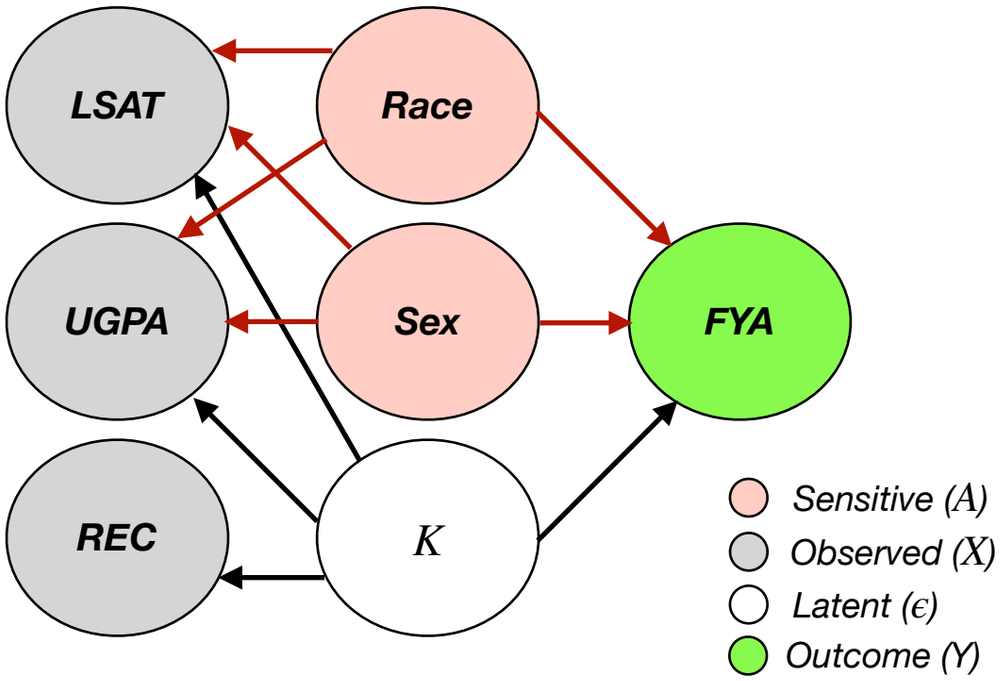
$$\textbf{REC} := \mathcal{N}(w_R^K K, 1)$$

$$K \sim \mathcal{N}(0,1)$$

Relations between the features

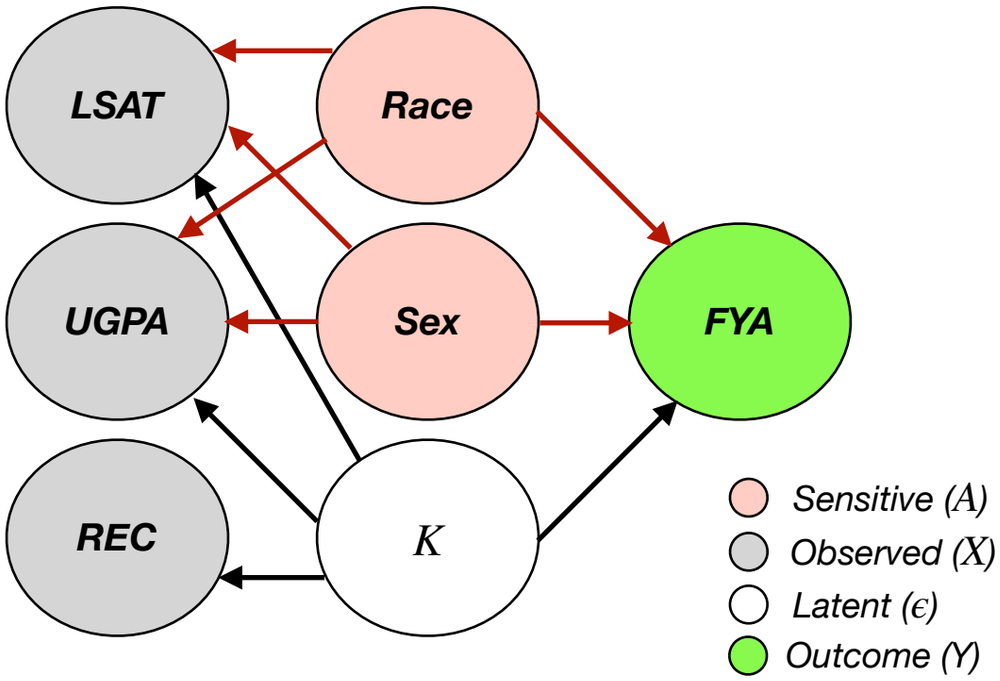
Quantification of the relations

Counterfactuals from causal model[1]



[1]*Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

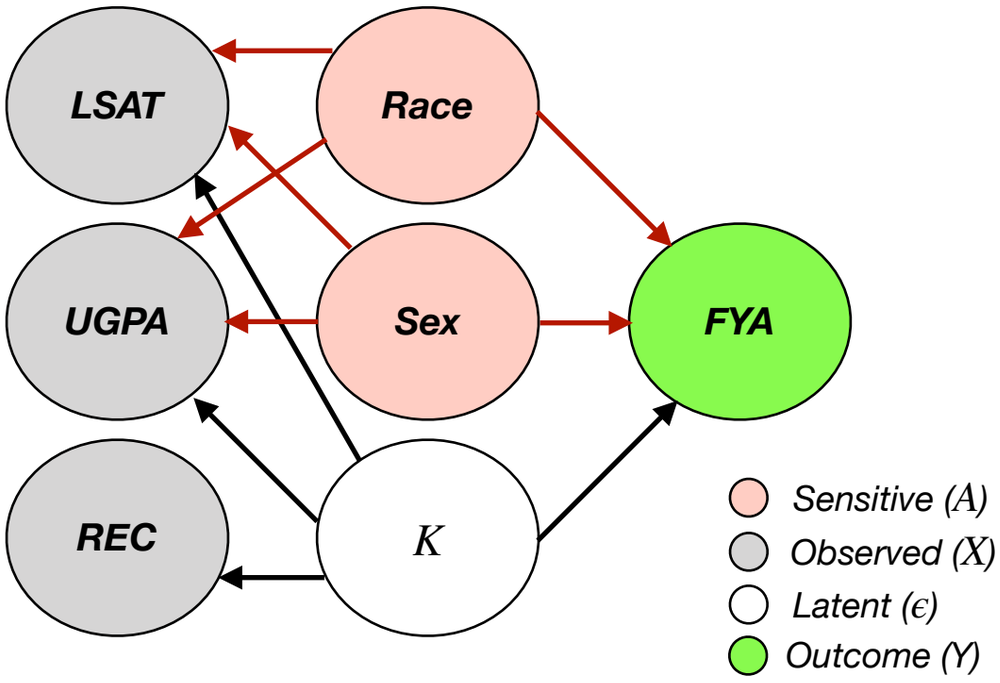
Counterfactuals from causal model[1]



1. *Abduction*: Given X, $A = a$ **estimate** ϵ

[1] *Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

Counterfactuals from causal model[1]

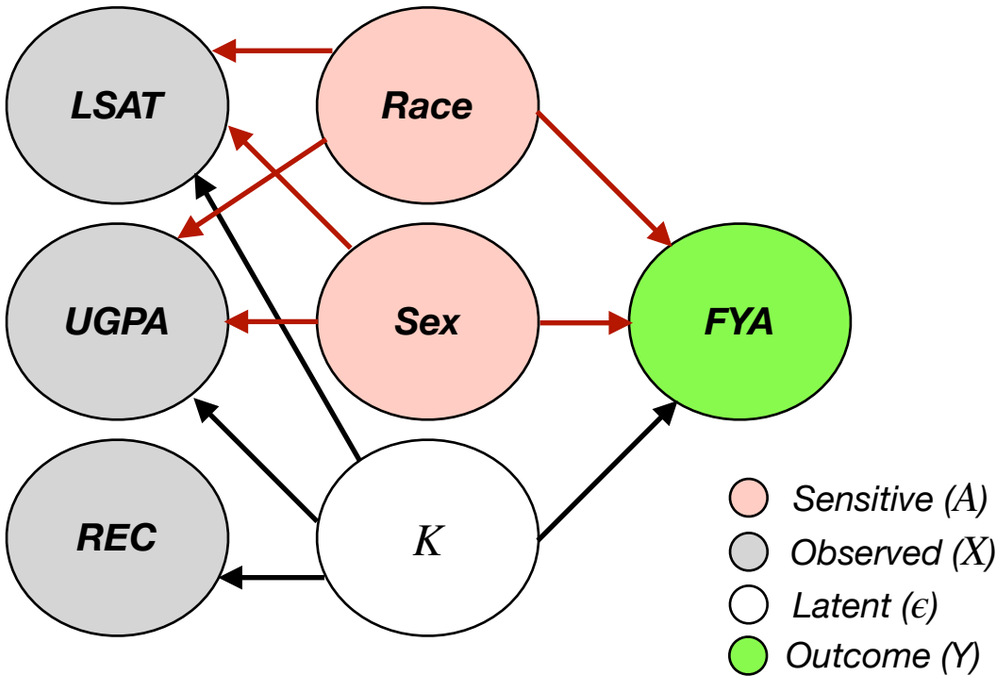


1. *Abduction*: Given X, $A = a$ **estimate** ϵ

2. *Action*: **Intervene** on A by setting it to a'

[1] *Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

Counterfactuals from causal model[1]



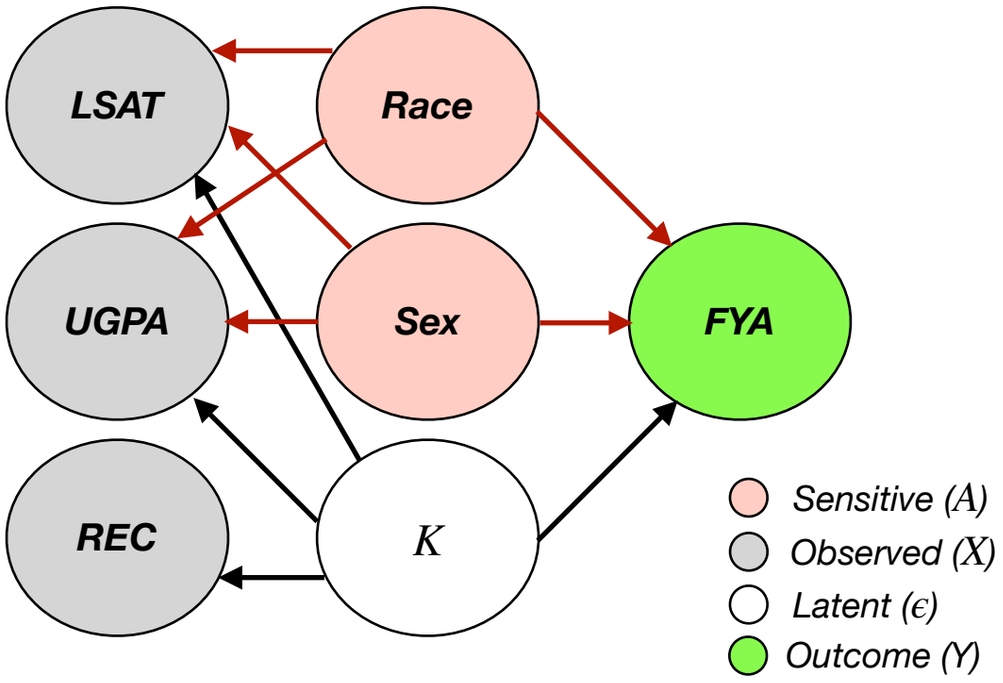
1. *Abduction*: Given X, $A = a$ **estimate** ϵ

2. *Action*: **Intervene** on A by setting it to a'

3. *Prediction*: **Re-compute** X using ϵ under intervention $do(A = a')$

[1] *Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

Counterfactuals from causal model[1]



1. *Abduction*: Given X, $A = a$ **estimate** ϵ

2. *Action*: **Intervene** on A by setting it to a'

3. *Prediction*: **Re-compute** X using ϵ under intervention $do(A = a')$

How to approximate in the absence of whole causal model?

[1] *Pearl, J. (2009). Causality: Models, reasoning, and inference, (2nd ed.). New York: Cambridge University Press.*

Assumptions

✗ **Problem:** No access to complete causal model

- Exact *causal graph* over all features **not known**

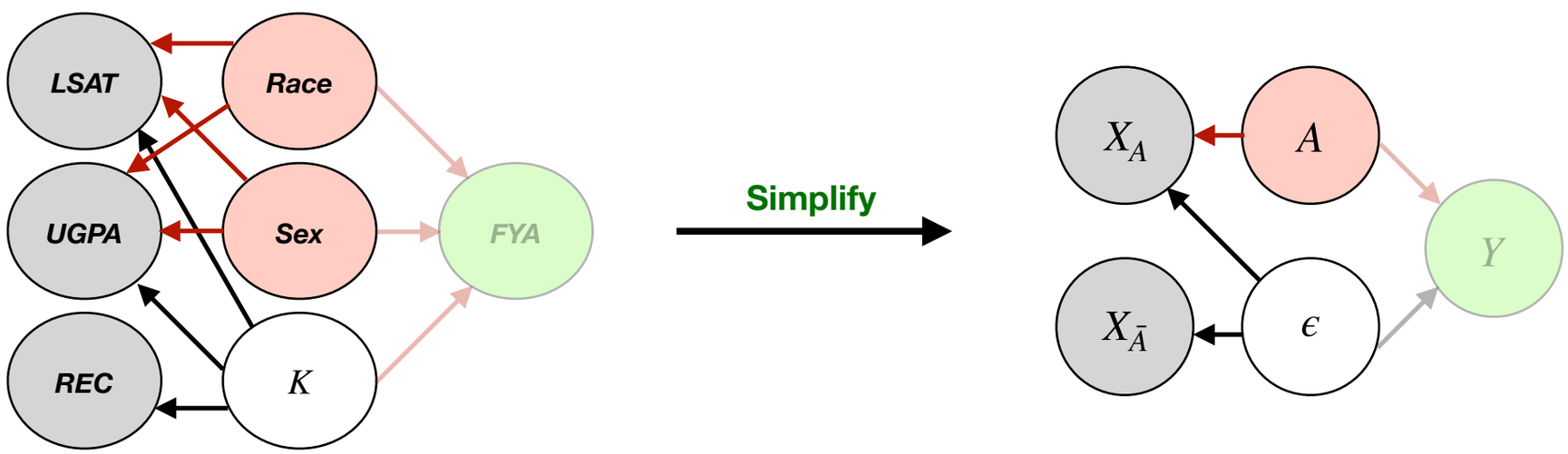
Assumptions

✗ **Problem:** No access to complete causal model

- Exact *causal graph* over all features **not known**

✓ **Simplify** assumptions on data generative process

- A are *root* nodes
- A can *affect* some features (X_A) or not ($X_{\bar{A}}$)
- Latent ϵ *affect* features X independent of A



Modeling

✗ **Problem:** No access to complete causal model

- Exact *structural equations* of all features **not known**

Modeling

✗ **Problem:** No access to complete causal model

- Exact *structural equations* of all features **not known**

✓ **Model** by approximating data generative process

- Use variational generative model *(Conditional VAE)*
- Learn generative process conditional on A
- Estimate latent factors of data via z

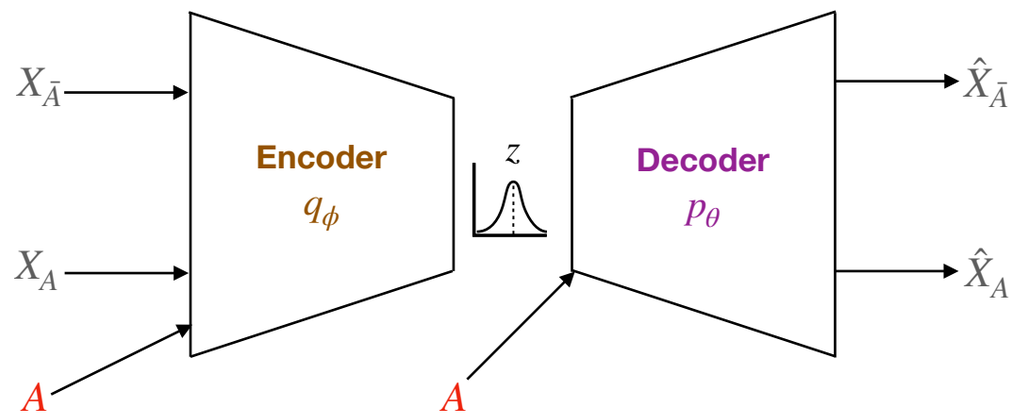
LSAT $\sim \ldots$ **UGPA** $\sim \ldots$

REC $\sim \ldots$ $\epsilon \sim \ldots$

Structural equations

Learn

from data



CVAE architecture

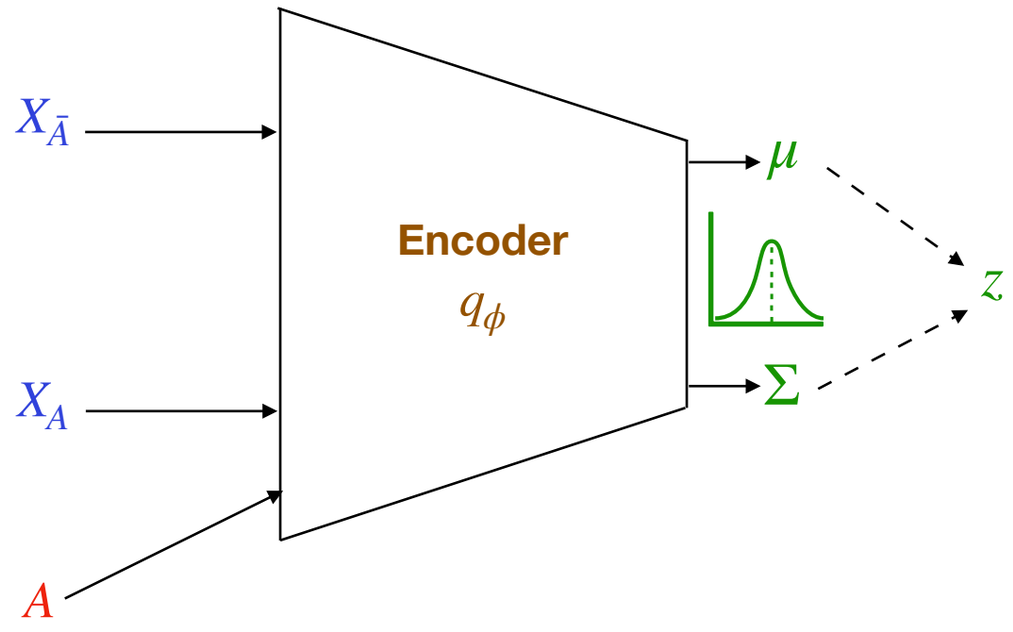
CVAE Model: Encoder

★ **Goal:** Learn latent distribution conditional on A given data

CVAE Model: Encoder

★ **Goal:** Learn latent distribution conditional on A given data

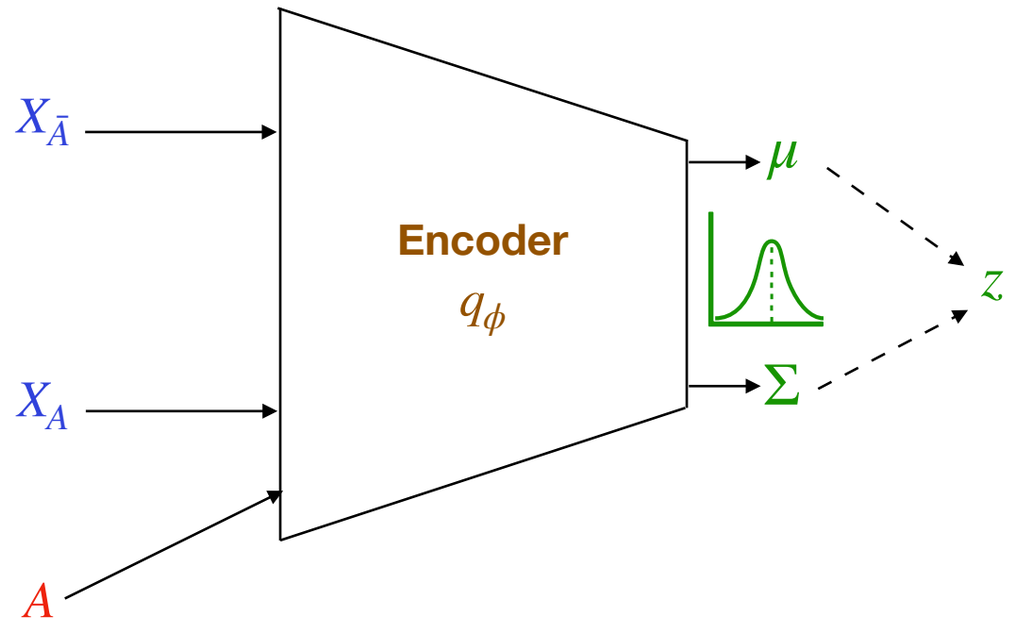
- Use **Encoder** of CVAE:
  - Deep neural net q with parameters ϕ
  - **Input**: *Data* $X_{\bar{A}}, X_A$; *Sensitive* A
  - **Output**: Latent z



CVAE Model: Encoder

★ **Goal:** Learn latent distribution conditional on A given data

- Use **Encoder** of CVAE:
  - Deep neural net q with parameters ϕ
  - **Input**: *Data* $X_{\bar{A}}, X_A$; *Sensitive* A
  - **Output**: Latent z
- z must match prior distribution
  - E.g. $p(z) \sim \mathcal{N}(0,1)$



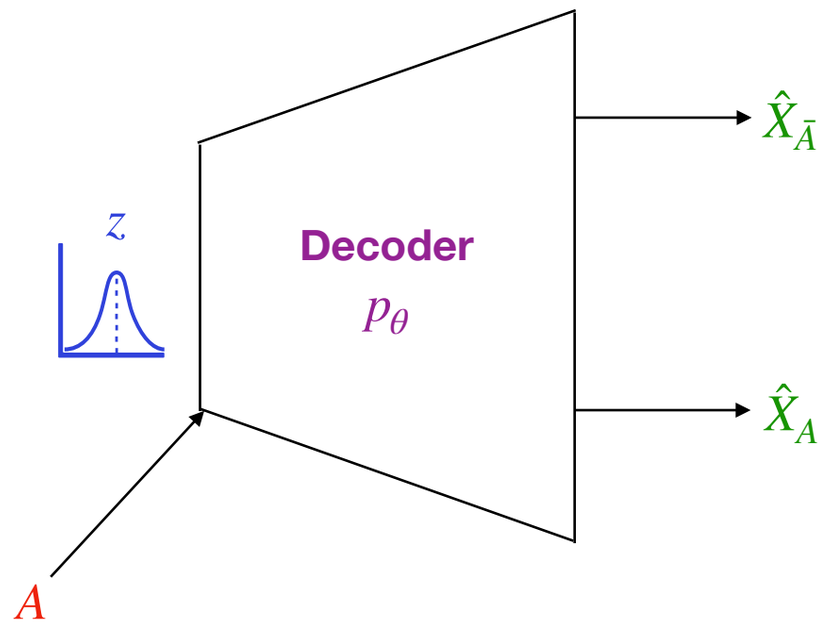
CVAE Model: Decoder

★**Goal:** Learn data distribution conditional on A given latent z

CVAE Model: Decoder

★**Goal:** Learn data distribution conditional on A given latent z

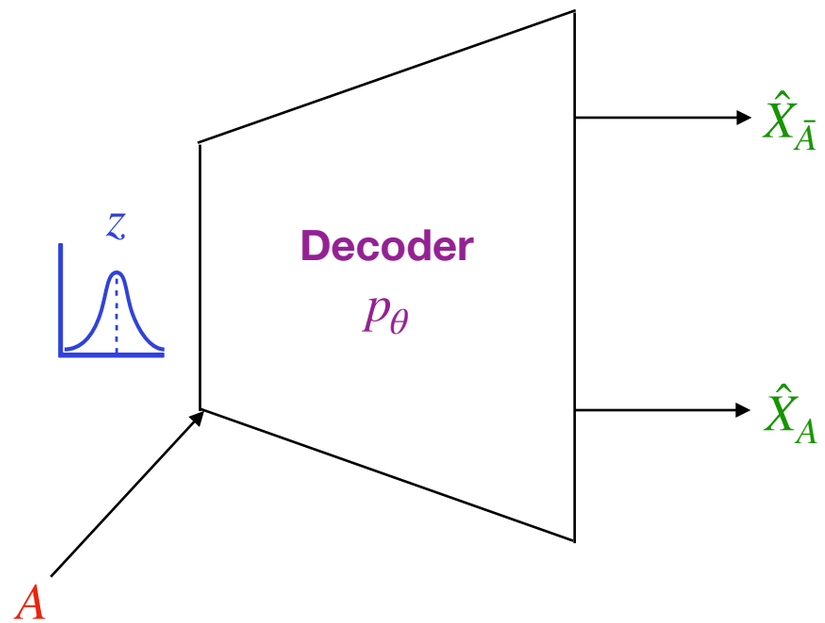
- Use **Decoder** of CVAE:
  - Deep neural net p with parameters θ
  - **Input**: *Latent factors* z ; *Sensitive* A
  - Output: Reconstructed data $\hat{X}_{\bar{A}}, \hat{X}_A$



CVAE Model: Decoder

★**Goal:** Learn data distribution conditional on A given latent z

- Use **Decoder** of CVAE:
  - Deep neural net p with parameters θ
  - **Input**: *Latent factors* z ; *Sensitive* A
  - Output: Reconstructed data $\hat{X}_{\bar{A}}, \hat{X}_A$



What do we optimize for?

Train models end-to-end

Evidence **L**ower **BO**und (**ELBO**) Loss

$$\log p_\theta(X \mid A) \geq \mathbb{E}_{q_\phi(z|X,A)}[\log p_\theta(X \mid z, A)] - \mathbb{D}_{KL}[q_\phi(z \mid X, A) \mid\mid p(z)]$$

Train models end-to-end

Evidence **L**ower **BO**und (**ELBO**) Loss

$$\log p_\theta(X \mid A) \geq \mathbb{E}_{q_\phi(z|X,A)}[\log p_\theta(X \mid z, A)] - \underbrace{\mathbb{D}_{KL}[q_\phi(z \mid X, A) \mid\mid p(z)]}$$

Encoder

- **Encoder**: z should match prior → *minimize KL* ***divergence***

Train models end-to-end

Evidence **L**ower **BO**und (**ELBO**) Loss

$$\log p_\theta(X \mid A) \geq \underbrace{\mathbb{E}_{q_\phi(z \mid X,A)}[\log p_\theta(X \mid z, A)]}_{\text{Decoder}} - \underbrace{\mathbb{D}_{KL}[q_\phi(z \mid X, A) \mid\mid p(z)]}_{\text{Encoder}}$$

- **Encoder**: z should match prior → *minimize KL* ***divergence***
- **Decoder**: Estimated z should maximize data-likelihood → *minimize reconstruction* ***loss***

CVAE Counterfactuals

Causal **CVAE**

CVAE Counterfactuals

Causal

CVAE

Infer ϵ from $X, A = a$

1. Abduction

Infer z from $X, A = a$

CVAE Counterfactuals

Causal		CVAE
Infer ϵ from $X, A = a$	**1. Abduction**	Infer z from $X, A = a$
Set $A = a'$ in the graph	**2. Action**	Set $A = a'$ at decoder

CVAE Counterfactuals

Causal		CVAE
Infer ϵ from $X, A = a$	**1. Abduction**	Infer z from $X, A = a$
Set $A = a'$ in the graph	**2. Action**	Set $A = a'$ at decoder
Get X from ϵ and $A = a'$	**3. Prediction**	Get X from z and $A = a'$

Counterfactual fairness

How to achieve fairness with counterfactuals?

Counterfactual fairness

How to achieve fairness with counterfactuals?

Prediction (for any individual) should not change while:

Counterfactual fairness

How to achieve fairness with counterfactuals?

Prediction (for any individual) should not change while:

- **Change** A
- Everything not **causally** dependent on A **constant**

Counterfactual fairness

How to achieve fairness with counterfactuals?

Prediction (for any individual) should not change while:

- **Change** A
- Everything not **causally** dependent on A **constant**

$$P\left(\hat{Y}_{A\leftarrow a}(\epsilon) = y \mid X = x, A = a\right) = P\left(\hat{Y}_{A\leftarrow a'}(\epsilon) = y \mid X = x, A = a\right)$$

Counterfactual fairness

How to achieve fairness with counterfactuals?

Prediction (for any individual) should not change while:

- **Change** A
- Everything not **causally** dependent on A **constant**

$$P\left(\hat{Y}_{A\leftarrow a}(\epsilon) = y \mid X = x, A = a\right) = P\left(\hat{Y}_{A\leftarrow a'}(\epsilon) = y \mid X = x, A = a\right)$$

A should not **cause** $\hat{Y}$ in any **individual** instance![1]

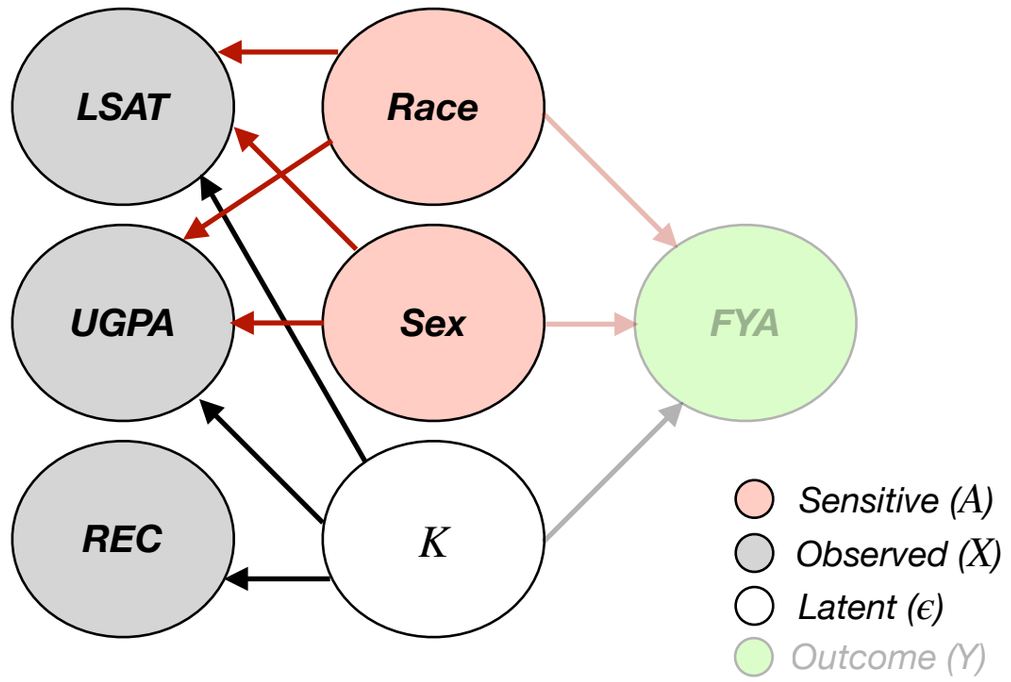
[1] *Matt J Kusner et al. "Counterfactual Fairness". In Advances in Neural Information Processing Systems 30.*

Results

How well can we approximate?

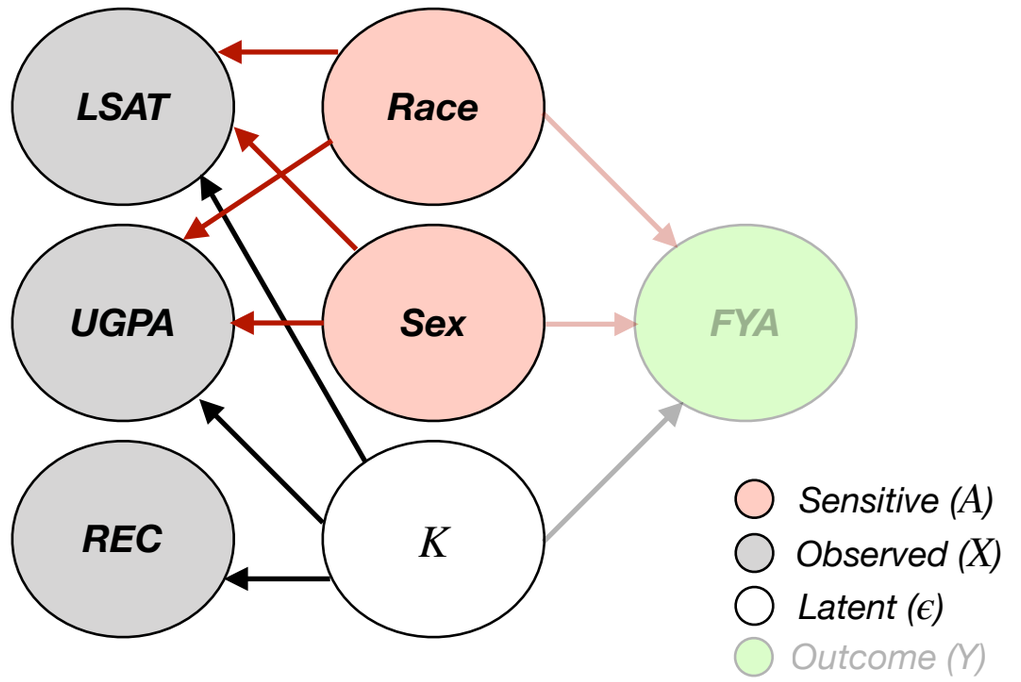
Generate counterfactuals

- **Train** CVAE on synthetic **generated data.**



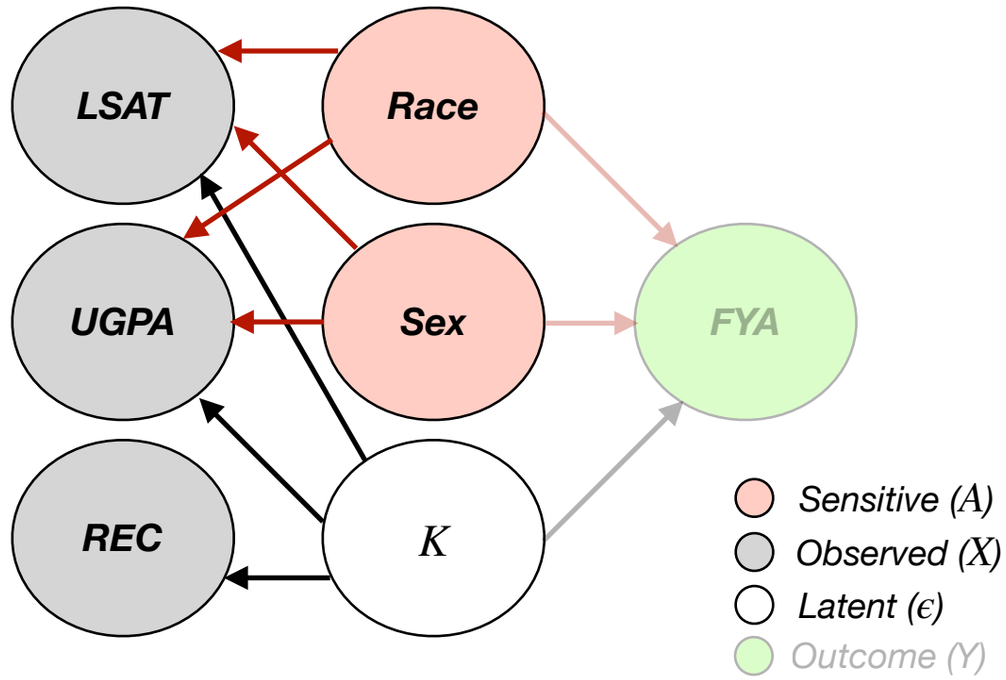
Generate counterfactuals

- **Train** CVAE on synthetic **generated data.**
- Condition on A **(race, sex)**.



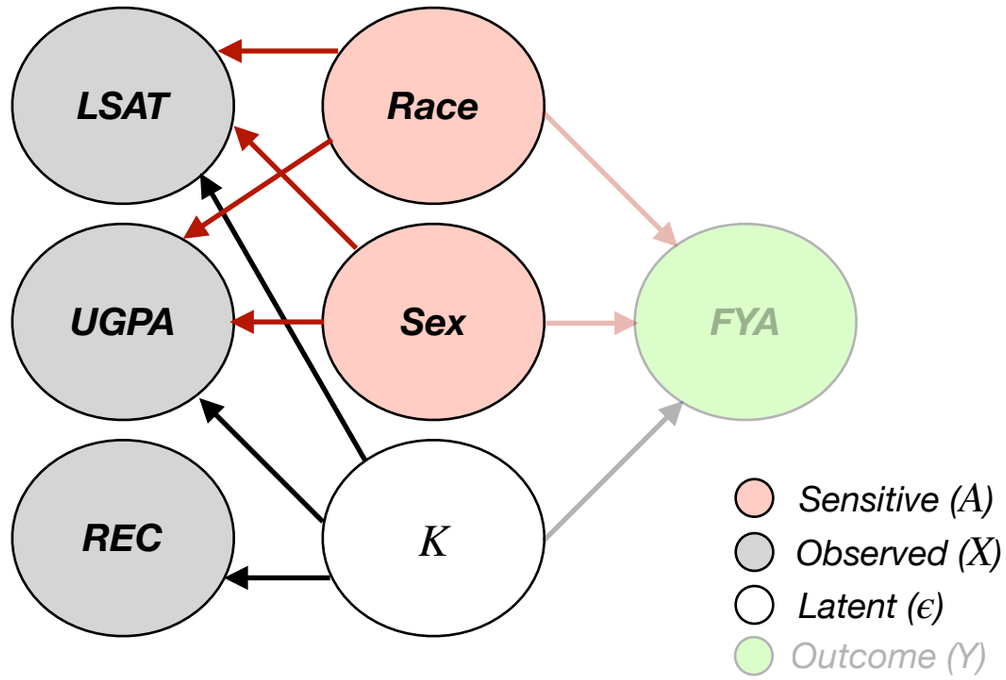
Generate counterfactuals

- **Train** CVAE on synthetic **generated data.**
- Condition on A **(race, sex)**.
- **Metrics:** **Mean absolute error** (Err.), **cosine similarity** (cos. sim.) b/w CVAE & causal counterfactuals.
  - Use standardized data



Generate counterfactuals

- **Train** CVAE on synthetic **generated data.**
- Condition on A **(race, sex)**.
- **Metrics:** **Mean absolute error** (Err.), **cosine similarity** (cos. sim.) b/w CVAE & causal counterfactuals.
  - Use standardized data

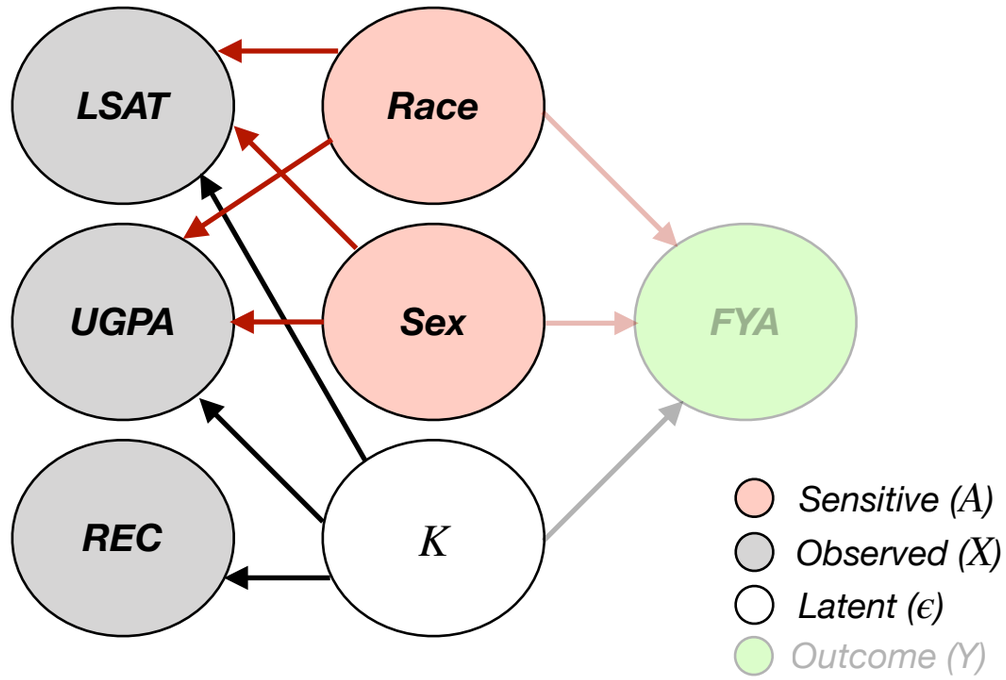


LSAT Err	UGPA Err	REC Err	Cos. Sim.
0.016	0.008	0.02	0.9998

Counterfactual generation quality (Race: White to Black).

Generate counterfactuals

- **Train** CVAE on synthetic **generated data.**
- Condition on A **(race, sex)**.
- **Metrics:** **Mean absolute error** (Err.), **cosine similarity** (cos. sim.) b/w CVAE & causal counterfactuals.
  - Use standardized data



LSAT Err	UGPA Err	REC Err	Cos. Sim.
0.016	0.008	0.02	0.9998

Counterfactual generation quality (Race: White to Black).

CVAE can generate faithful counterfactuals!

Can we use generated **counterfactuals** for auditing?

Auditing setup

Auditing setup

- **Trained** regression model
  - Predict FYA score (standardized)
  - Audit w.r.t. race (White-Black)

Auditing setup

- **Trained** regression model
  - Predict FYA score (standardized)
  - Audit w.r.t. race (White-Black)
- Audit counterfactual fairness:

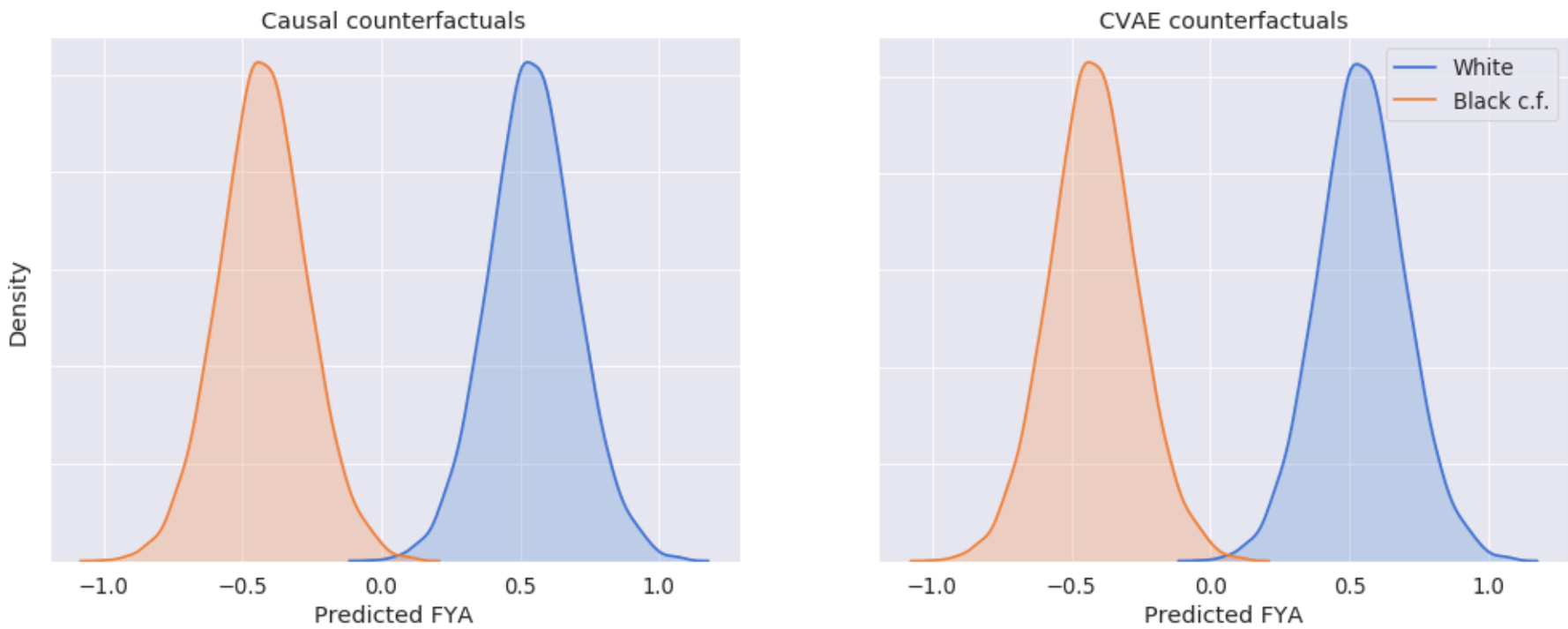
Auditing setup

- **Trained** regression model
  - Predict FYA score (standardized)
  - Audit w.r.t. race (White-Black)
- Audit counterfactual fairness:
  - ***White*** *individual was predicted to have standardized FYA score 1.*
  - *If individual was* ***black*** *instead, would the* ***predicted score change****?*

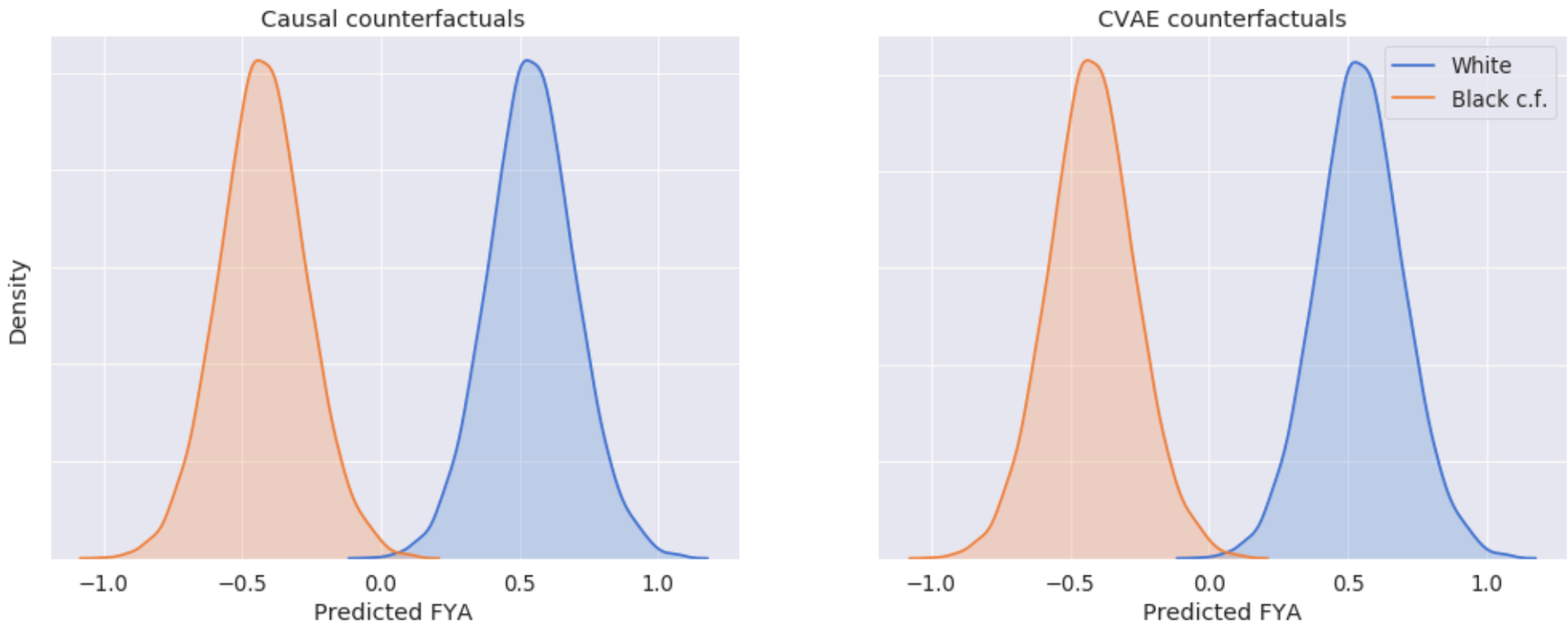
Auditing setup

- **Trained** regression model
  - Predict FYA score (standardized)
  - Audit w.r.t. race (White-Black)
- Audit counterfactual fairness:
  - ***White*** *individual was predicted to have standardized FYA score 1.*
  - *If individual was* ***black*** *instead, would the* ***predicted score change****?*
- CVAE generated counterfactuals to audit model
  - Compare with true causal auditing

Audit counterfactual fairness



Audit counterfactual fairness



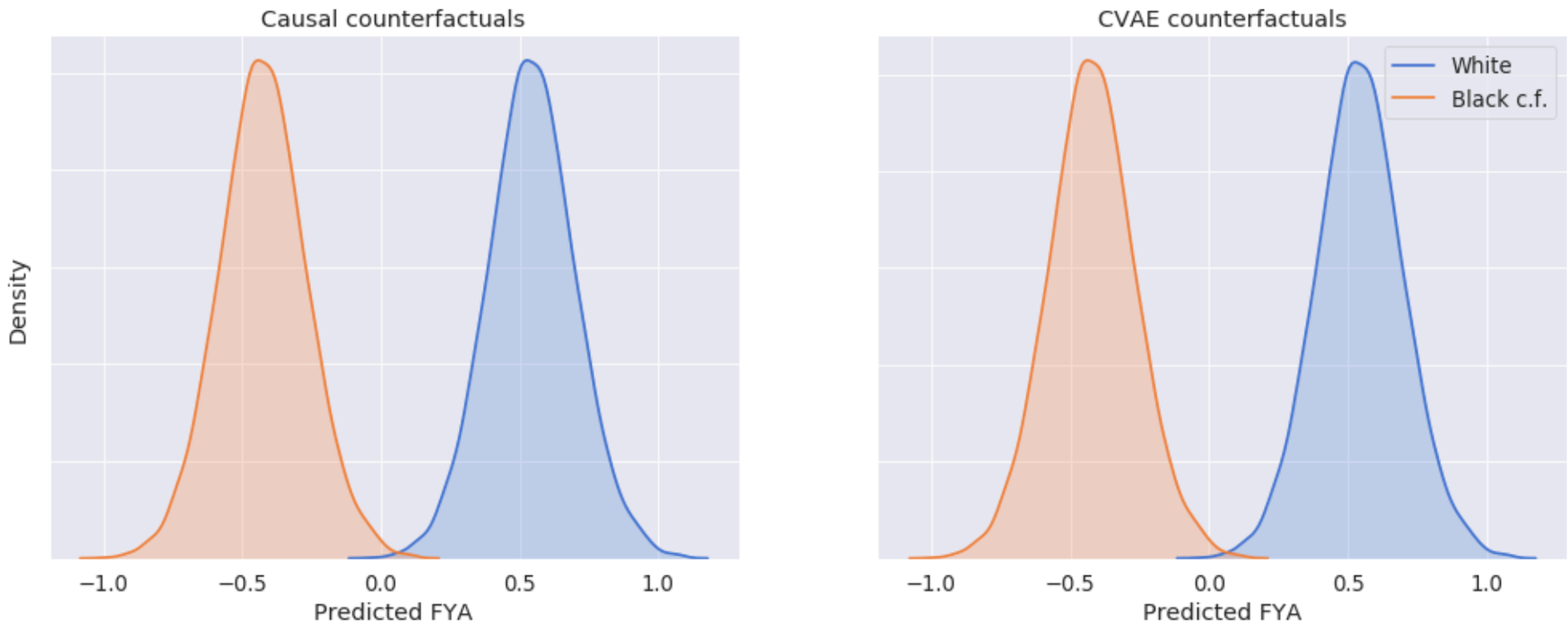
White → Black

Change in predicted scores:

Causal: ***0.97*** ↓ *; CVAE:* ***0.96*** ↓

White → Black : : Predicted score reduces!

Audit counterfactual fairness



White → Black

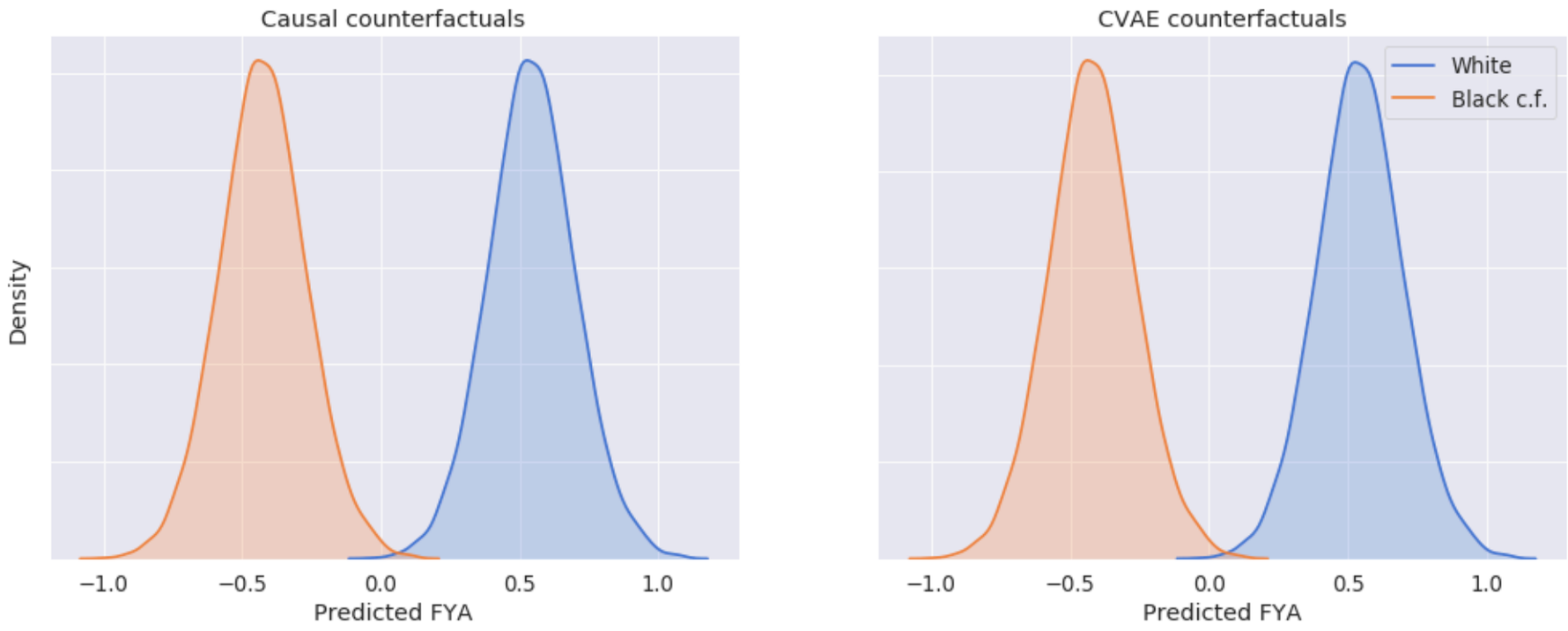
Change in predicted scores:

Causal: ***0.97*** ↓ *; CVAE:* ***0.96*** ↓

White → Black : : Predicted score reduces!

Model biased negatively towards blacks!

Audit counterfactual fairness



White → Black

Change in predicted scores:

Causal: ***0.97*** ↓ *; CVAE:* ***0.96*** ↓

White → Black : : Predicted score reduces!

Model biased negatively towards blacks!

CVAE auditing ≃ True causal auditing

Can we train a **fair** predictive system using our model?

Fair predictor setup

★ **Goal:** Train fair regression model to predict FYA

Fair predictor setup

★ **Goal:** Train fair regression model to predict FYA

Compare following models:

- **Aware**: Use all data features (incl. A)

Fair predictor setup

★ **Goal:** Train fair regression model to predict FYA

Compare following models:

- **Aware**: Use all data features (incl. A)
- **Unaware**: Use all features except A

Fair predictor setup

★ **Goal:** Train fair regression model to predict FYA

Compare following models:

- **Aware**: Use all data features (incl. A)
- **Unaware**: Use all features except A
- **Fair-z**: Train on CVAE latent z given data

Fair predictor setup

★ **Goal:** Train fair regression model to predict FYA

Compare following models:

- **Aware**: Use all data features (incl. A)
- **Unaware**: Use all features except A
- **Fair-z**: Train on CVAE latent z given data

Metrics:

- **Accuracy**: Root mean squared error (RMSE)
- **Unfairness**: Absolute difference in outcome to counterfactual

Fair predictor setup

★ **Goal:** Train fair regression model to predict FYA

Compare following models:

- **Aware**: Use all data features (incl. A)
- **Unaware**: Use all features except A
- **Fair-z**: Train on CVAE latent z given data

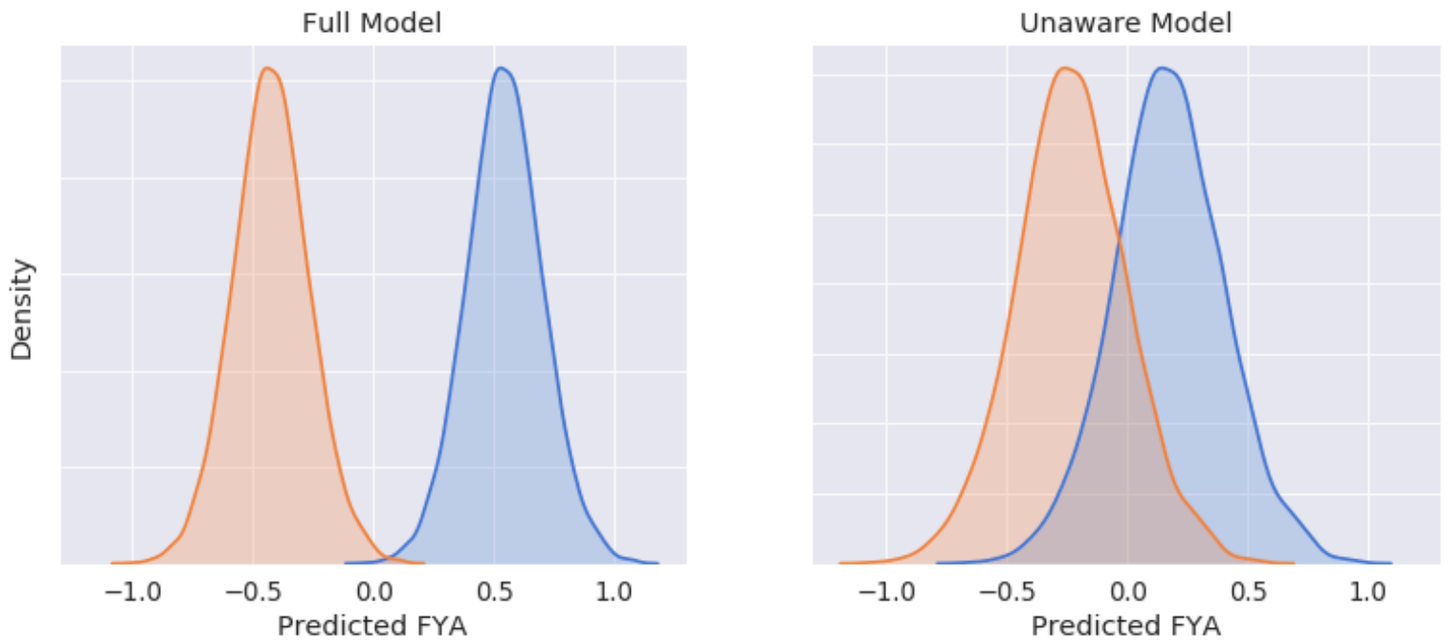
Metrics:

- **Accuracy**: Root mean squared error (RMSE)
- **Unfairness**: Absolute difference in outcome to counterfactual

Use data and its **causal counterfactual** for testing

Training fair predictor

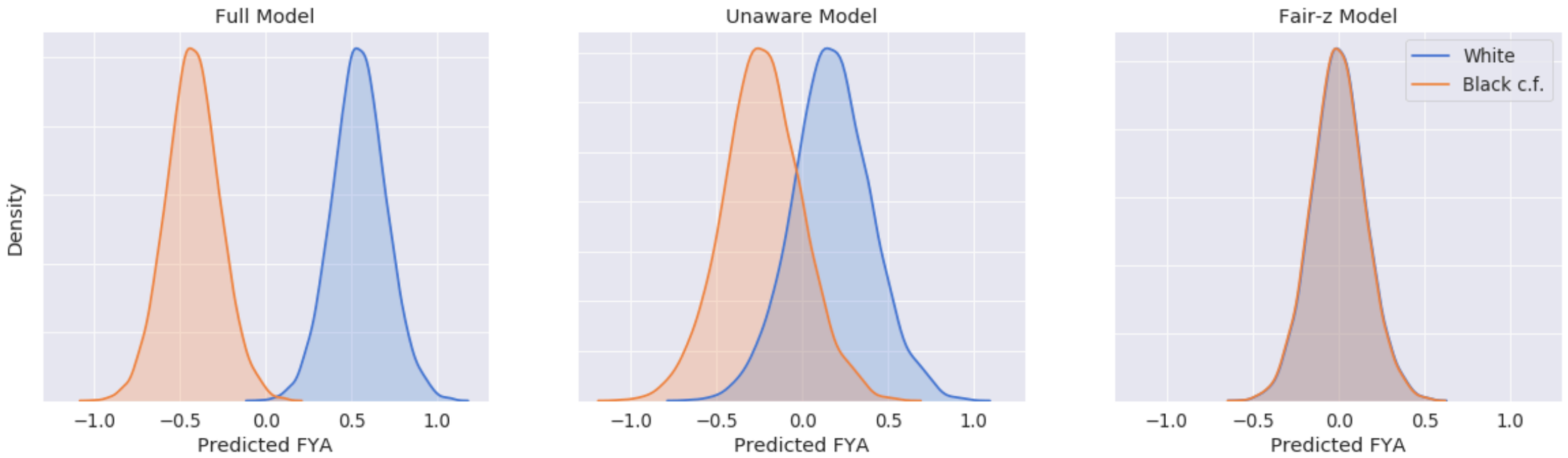
Training fair predictor



Distribution of predicted FYA output (White and counterfactual Black)

Model	Pred. Error (*RMSE*)	Unfairness (*Abs. Diff.*)
Full	**0.94 (very accurate)**	**0.97 (highly biased)**
Unaware	**1.0 (accurate)**	**0.40 (less biased)**

Training fair predictor



Distribution of predicted FYA output (White and counterfactual Black)

Model	Pred. Error (*RMSE*)	Unfairness (*Abs. Diff.*)
Full	**0.94 (very accurate)**	**0.97 (highly biased)**
Unaware	**1.0 (accurate)**	**0.40 (less biased)**
Fair-z	**1.07 (less accurate)**	**0.003 (fair)**

Related Work

FlipTest[1]

- Use GANs to approximate optimal transport
- No causal assumptions

Variational Fair Autoencoders[2]

- Representation learning to achieve demographic parity
- Use outcome label to encode representation

Fairness through Causal Awareness[3]

- VAE in causal setup
- Learns fair treatments w.r.t. outcome

[1] *Emily Black et al. “FlipTest: fairness testing via optimal transport”: 2020 Conference on Fairness, Accountability and Transparency*
[2] *Christos Louizos et al. "The Variational Fair Autoencoder": 2016 International Conference on Learning Representations*
[3] *David Madras et al. "Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data": 2019 Conference on Fairness, Accountability and Transparency*

Conclusions

- Causal analysis **vital** tool for fairness: use counterfactuals
- Used CVAE to generate counterfactuals under **minimal** causal assumptions
- Applied system to **audit** existing models for counterfactual fairness
- Used **latent** features extracted by system to train fair prediction model

Next steps

- In-depth comparison of our system with related methods
- Relate counterfactual and other notions of fairness w.r.t. our system
- Analyze limitations of our system w.r.t. counterfactual generation
- Explore applications in the real-world setting

Thank you!