Causality for ML Fairness

Sami Zhioua

March 9th, 2023

TAU Seminar

Fairness-Causality sub-team at Comète

Catuscia Palamidessi
Director of research at Inria and leader of Comète team
catuscia@lix.polytechnique.fr

Frank Valencia
CNRS Researcher
frank.valencia@inria.fr

Sami Zhioua
Advanced researcher at Inria, LIX, École Polytechnique
sami.zhioua@lix.polytechnique.fr

Ruta Binkyte
PhD student at Inria, LIX, École Polytechnique
ruta.binkyte-sadauskiene@inria.fr

Mario Alvim
Researcher
mario.ferreira-alvim-junior@inria.fr>

Karima Makhlouf
PhD student at Inria, LIX, École Polytechnique
karima.makhlouf@lix.polytechnique.fr

Carlos Pinzón
PhD student at Inria, LIX, École Polytechnique
carlos.pinzon@inria.fr

Héber H. Arcolezi
Posdoctoral Researcher at Inria, LIX, École Polytechnique
heber.hwang-arcolezi@inria.fr

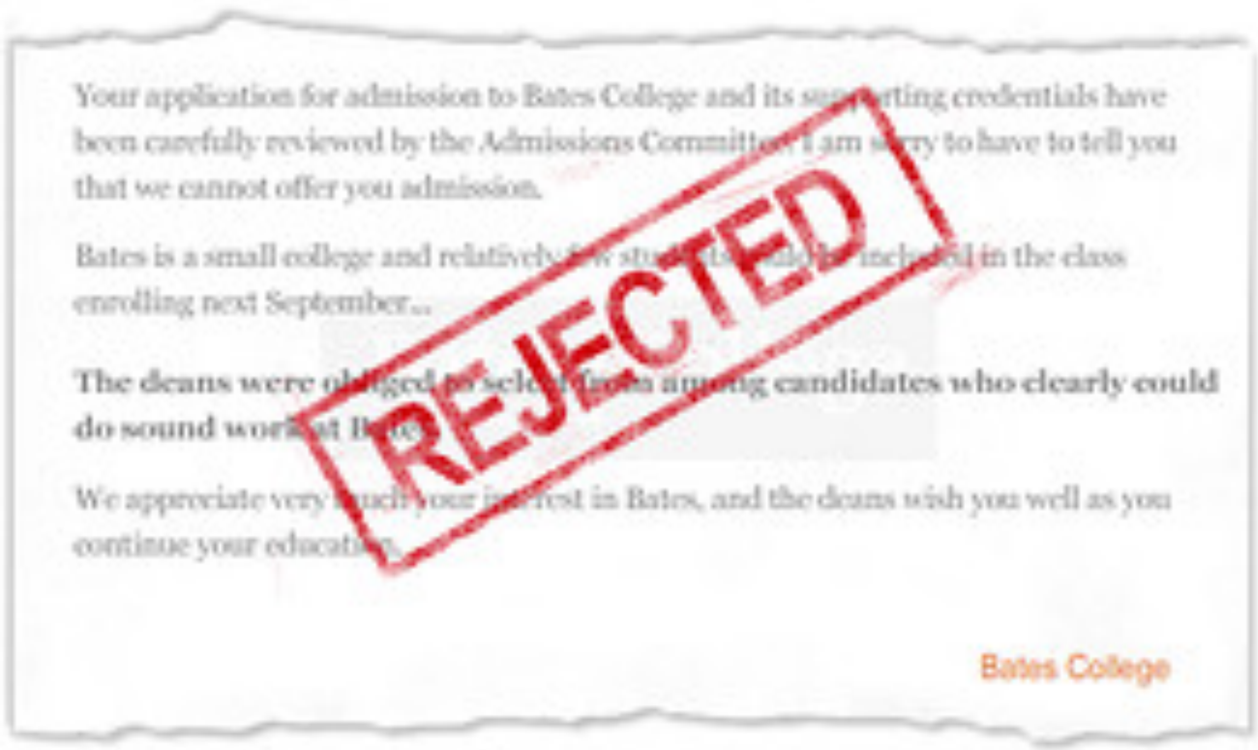
Szilvia Lestyan
Postdoctoral researcher

ML APPLICATIONS

Candidate Selection for Job Hiring

entelo

University Admission



predicting whether released people from jail will re-offend

COMPAS

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

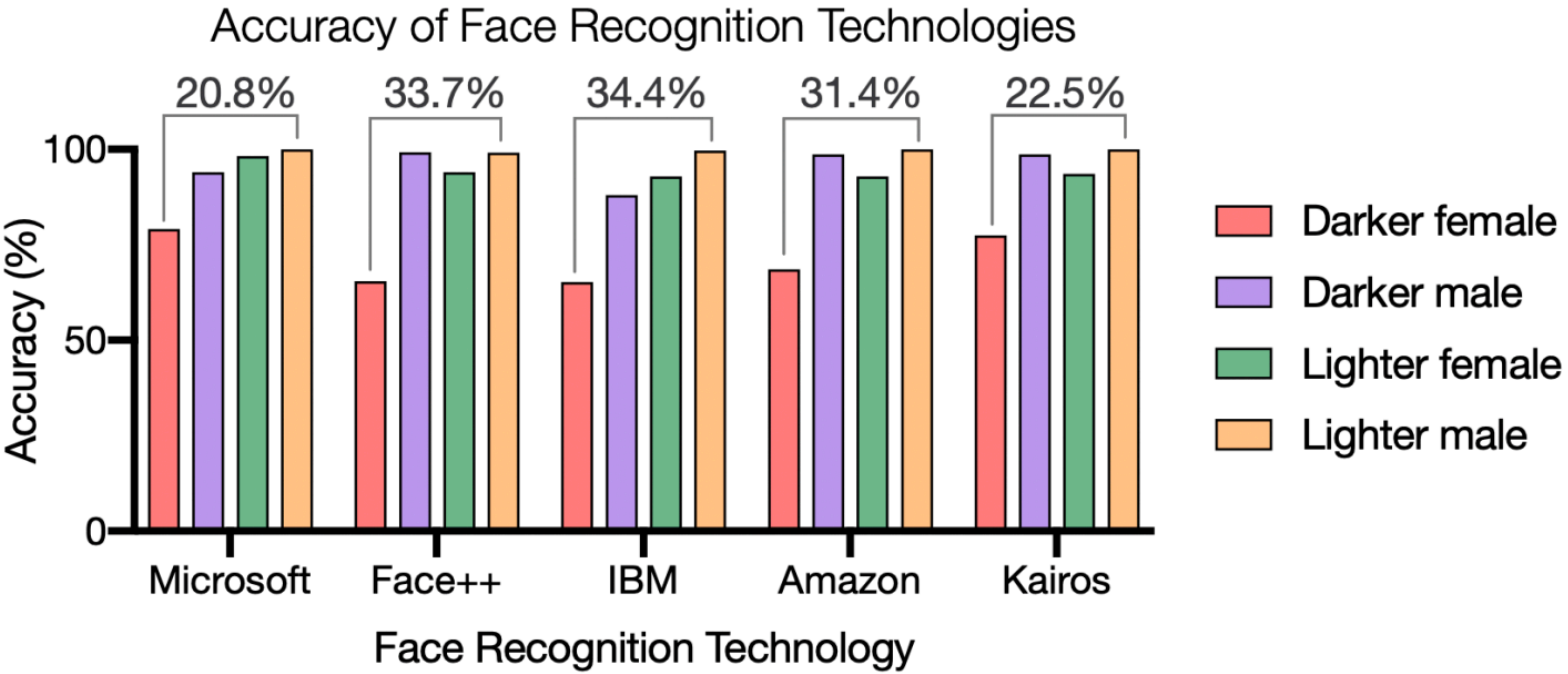
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

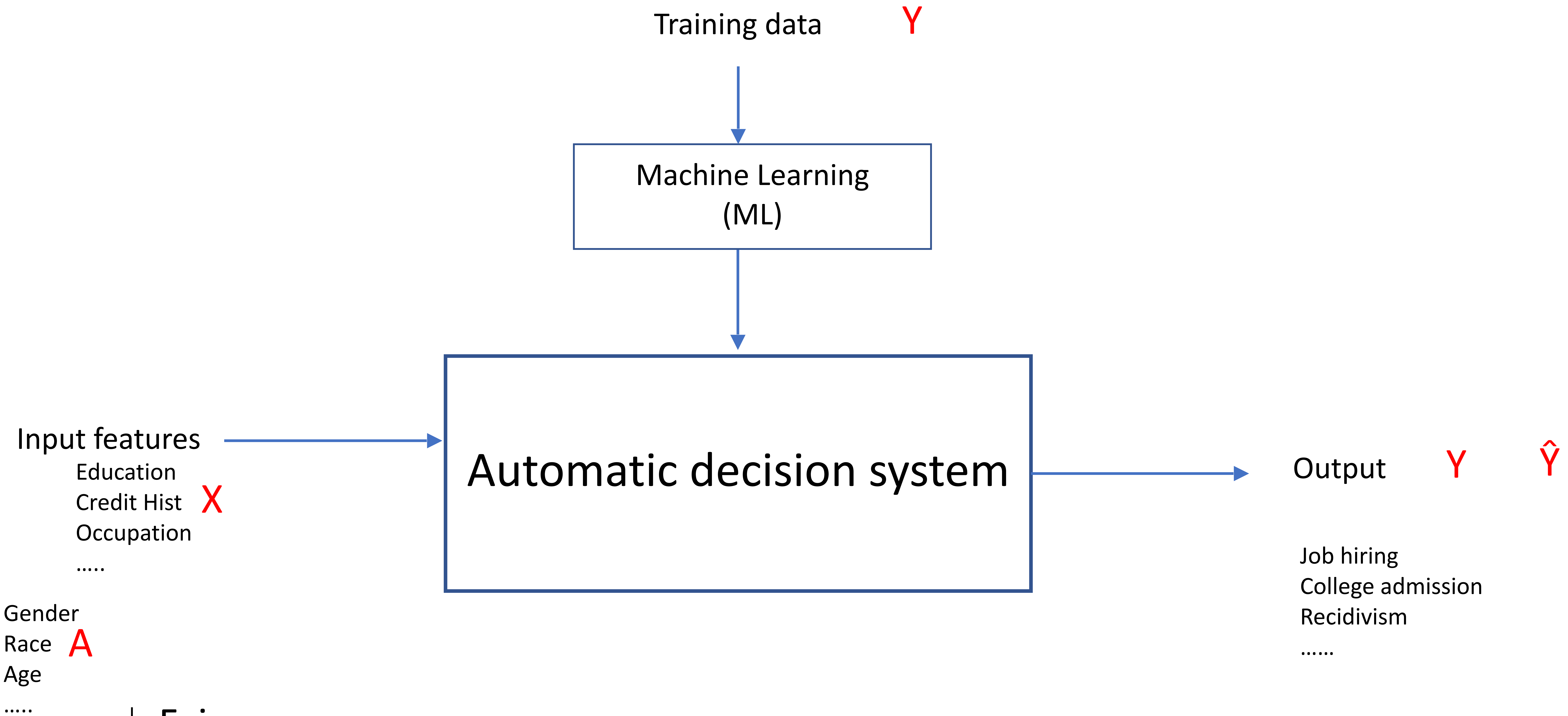
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.



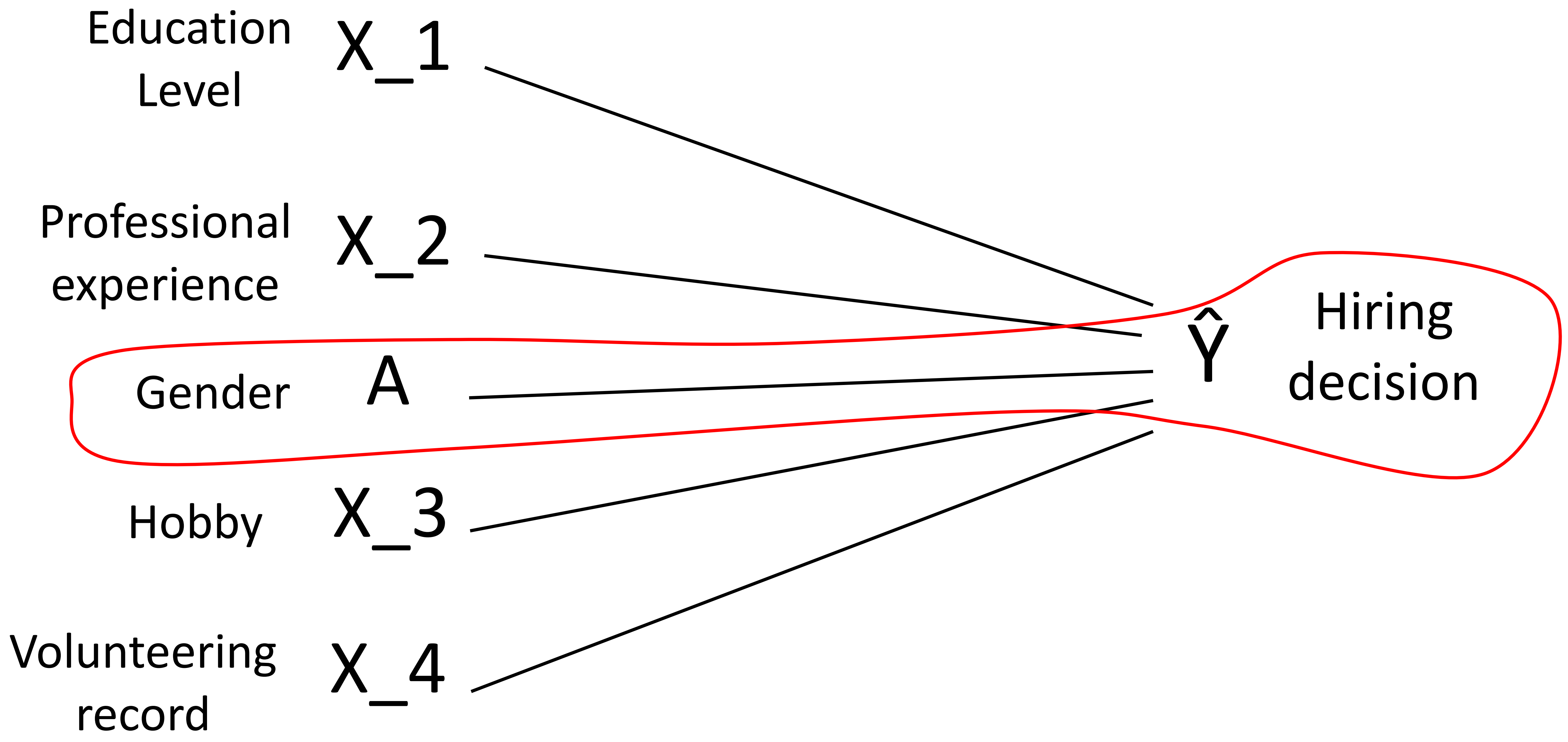
Ethical Concerns

Fairness: Is the output fair with respect to individuals or subpopulations ?

Explainability: How the output can be explained in terms of the input features ?

Privacy: Does learning high accuracy/utility model reveal personal and highly sensitive data?

Statistical notions of fairness



Female Male

$$P(\hat{Y} \mid A = 0) = P(\hat{Y} \mid A = 1)$$

Statistical Parity

$$P(\hat{Y} = 1 \mid E = e, A = 0) = P(\hat{Y} = 1 \mid E = e, A = 1) \quad \forall e$$

Conditional Statistical Parity

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1)$$

Equal Opportunity

(Equal TPRs)

$$E[S \mid Y = 1, A = 0)] = E[S \mid Y = 1, A = 1]$$

Balance

$$P(Y = 1 \mid \hat{Y} = 1, A = 0) = P(Y = 1 \mid \hat{Y} = 1, A = 1)$$

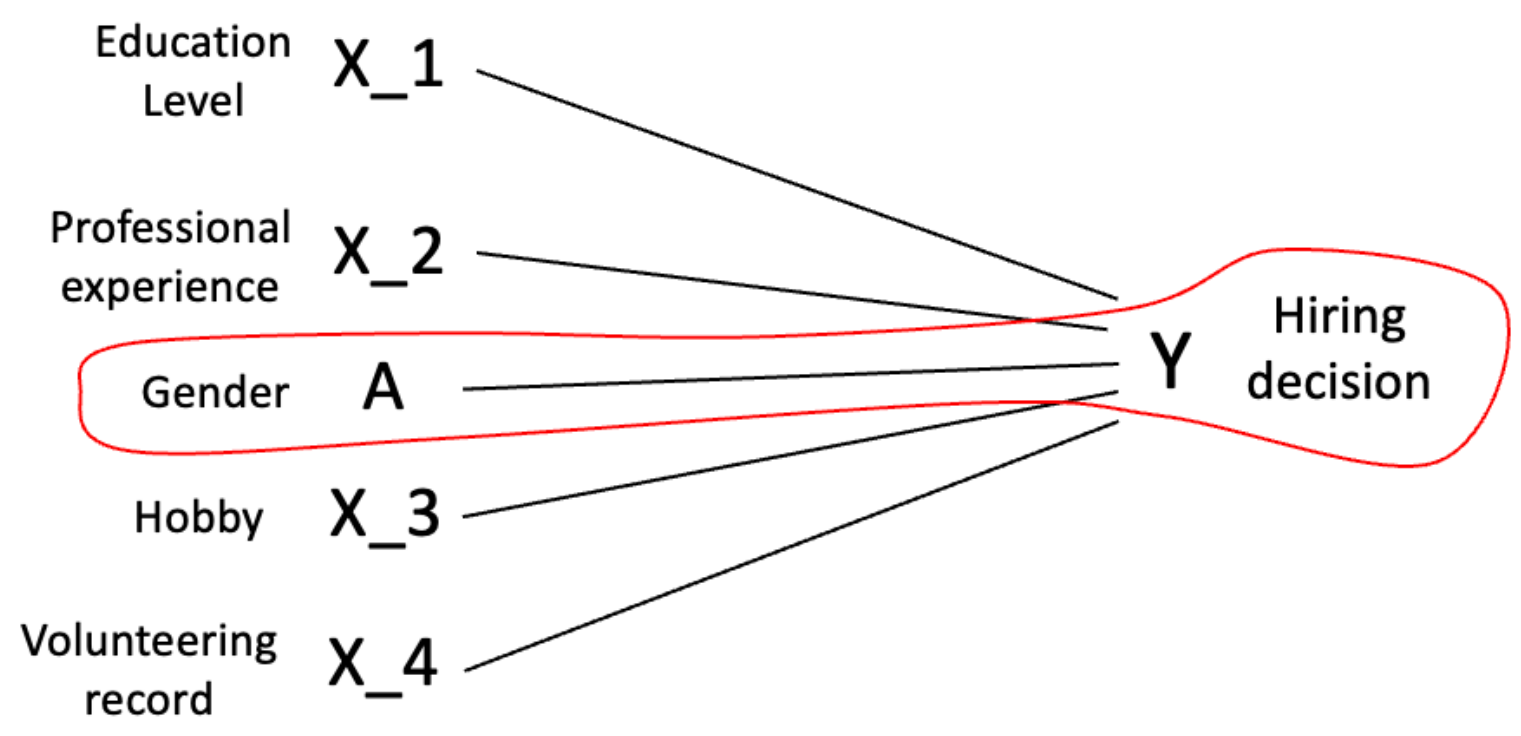
Predictive Parity

(Equal PPVs (Positive Predictive Values))

$$P(Y = 1 \mid S = s, A = 0) = P(Y = 1 \mid S = s, A = 1) \quad \forall s \in [0, 1]$$

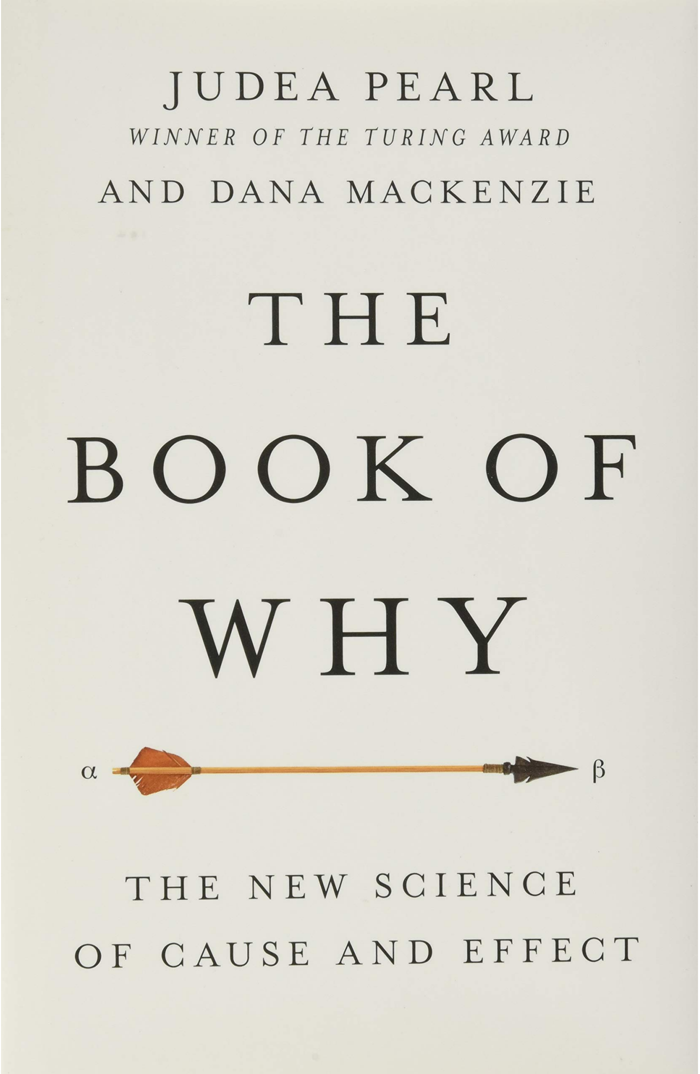
Calibration

How strong is the effect of A on Y ?



Why not P(Y|A)? → Bias

The illusion of correlation



"The correlation we observe is an illusion. An illusion we brought upon ourselves by choosing which events to include in our dataset and which to ignore."

Example 1:

Flip two coins 100 times, and write down the results only when at least one of them comes up head

Coin 1	Coin 2
Head	head
Tail	head
head	tail
Tail	head
Head	head

Notice the dependence: every time coin1 lands tail, coin2 lands head !

Example 2:

Did you notice that among the people you date, the attractive ones are more likely to be jerks ?

You are dating from these:

- Attractive Jerk
- Attractive Nice
- Not attractive Nice

Not attractive Jerk

Simpson's Paradox

Statistical parity = 7/15 – 8/15 = **-1/15**

Discrimination <u>against</u> women

Discrimation <u>in favor </u>of women

Hiring rate (T = 0) 3/10 = 0.3

Hiring rate (T = 1) 4/5 = 0.8

Total hiring rate 7/15

	A	T	Ŷ
	Gender	Job Type	Hiring
A=1 (Women)	1	0	1
	1	0	1
	1	0	1
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	1	1
	1	1	1
	1	1	1
	1	1	1
	1	1	0

	A	T	Ŷ
	Gender	Job Type	Hiring
A=0 (Men)	0	0	1
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	0
	0	1	0
	0	1	0

Hiring rate (T=0) 1/5 = 0.2

Hiring rate (T=1) 7/10 = 0.7

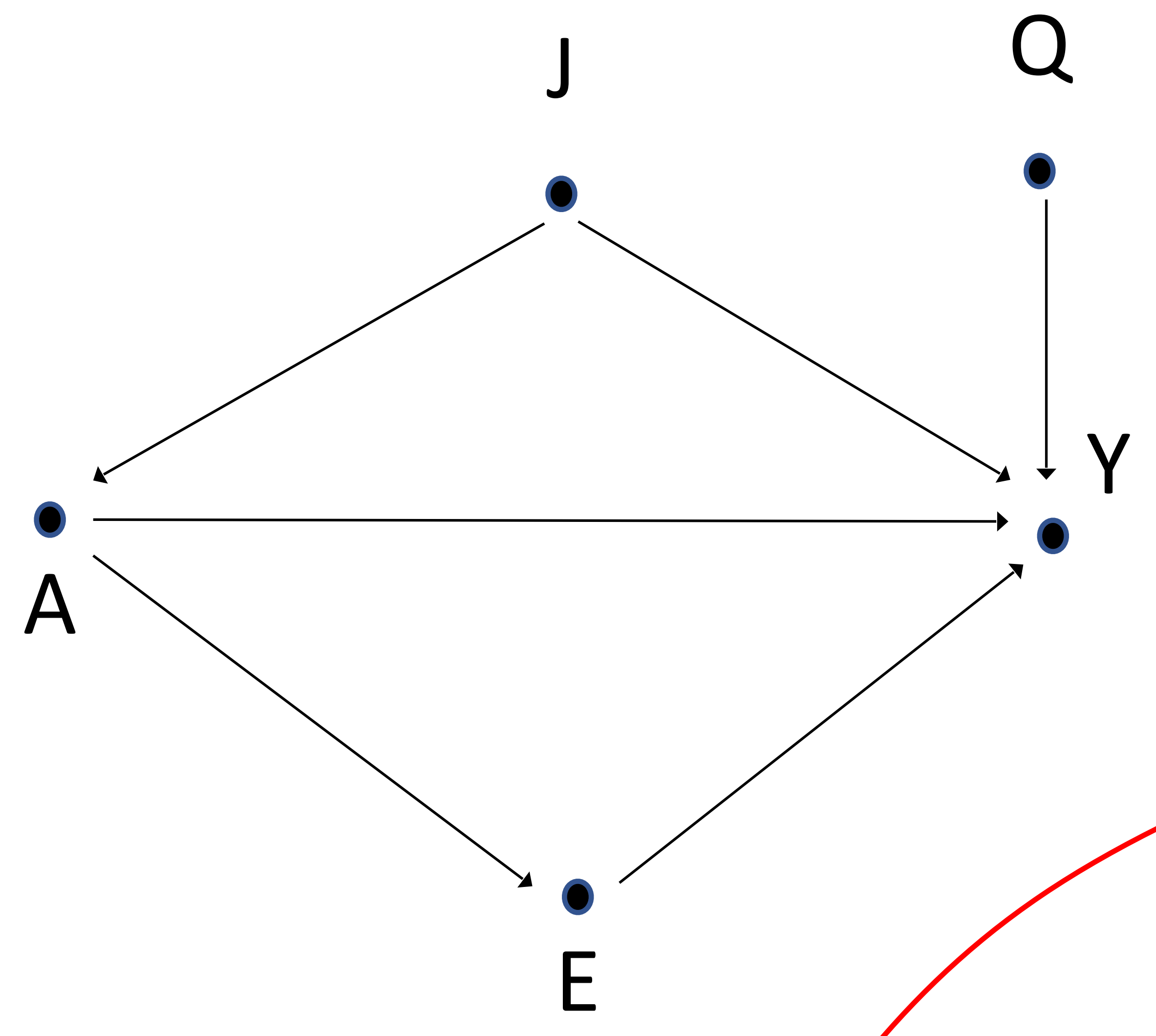
Total hiring rate 8/15

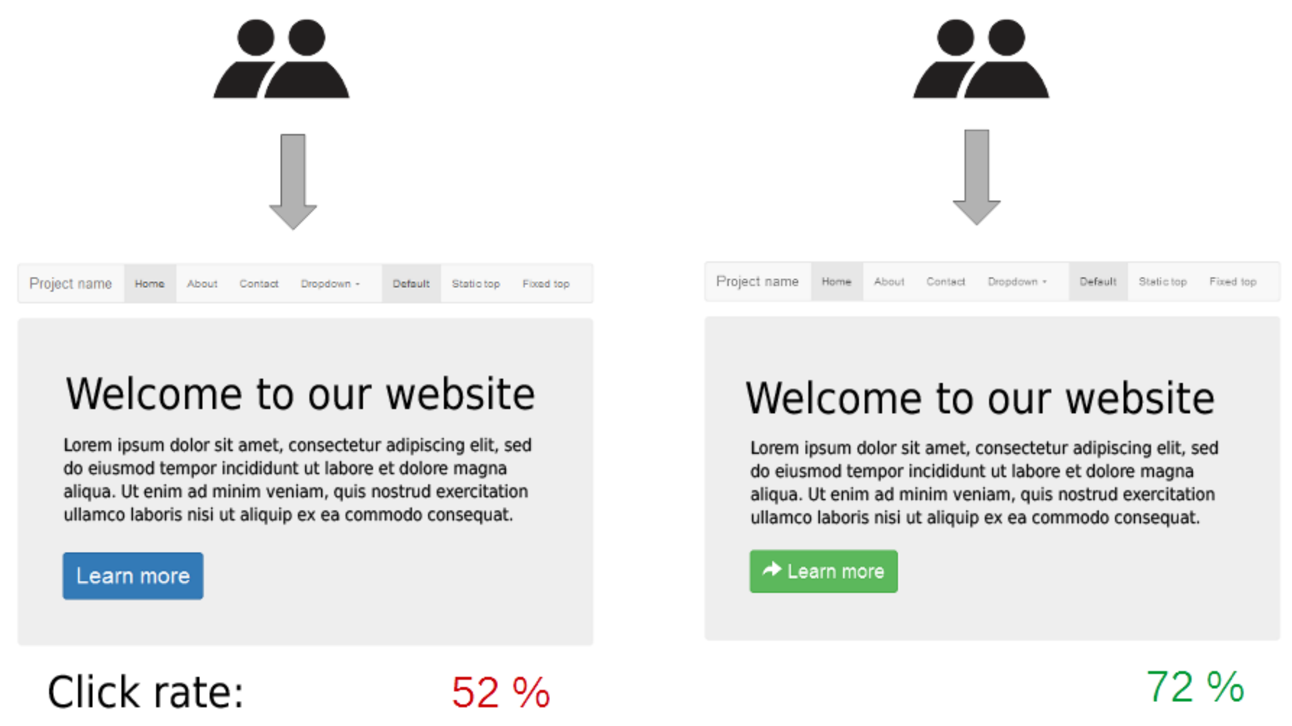
A = 0	Man
A = 1	Woman

T = 0	Flexible time job
T = 1	Non-flexible time job

Y=0	Not hired
Y=1	Hired

How to measure the causal effect reliably ?





A/B testing

The golden standard to measure causal effects is:

Randomized Controlled Trials (RCT)

Randomly allocating subjects to two or more groups

Treatment
(Receives the intervention)

Comparison

Control
(No-intervention, Placebo, etc.)

- It is the experimenter that does the allocation (not the subjects that choose)
- The experiment should be properly randomized:

All factors that influence the outcome variable are either static, or vary at random, except one

$\Rightarrow$ So any change in the outcome variable must be due to that one input variable.

An experiment involves an action (not mere observation)

In medical studies: select half of individuals randomly, and give them the treatment

In fairness problems: select half of candidates and ***set*** their gender to protected group (female).

How to measure the causal effect reliably ?

Causal Inference

Intervention: setting the value of a variable do(A = a)

P(Y=y|**A=a**)

The population distribution of Y among individuals whose A value is a

P(Y=y|**do(A=a)**)

The population distribution of Y if everyone in the population had their A value fixed at a.

Statistical Parity (Total Variation):

P(Y=y|A=1) - P(Y=y|A=0)

Total (causal) Effect:

TE = ATE = ACE =

P(Y=1|do(A=1)) – P(Y=1|do(A=0))

Other notations of P(Y=y|do(A=a)) in the literature

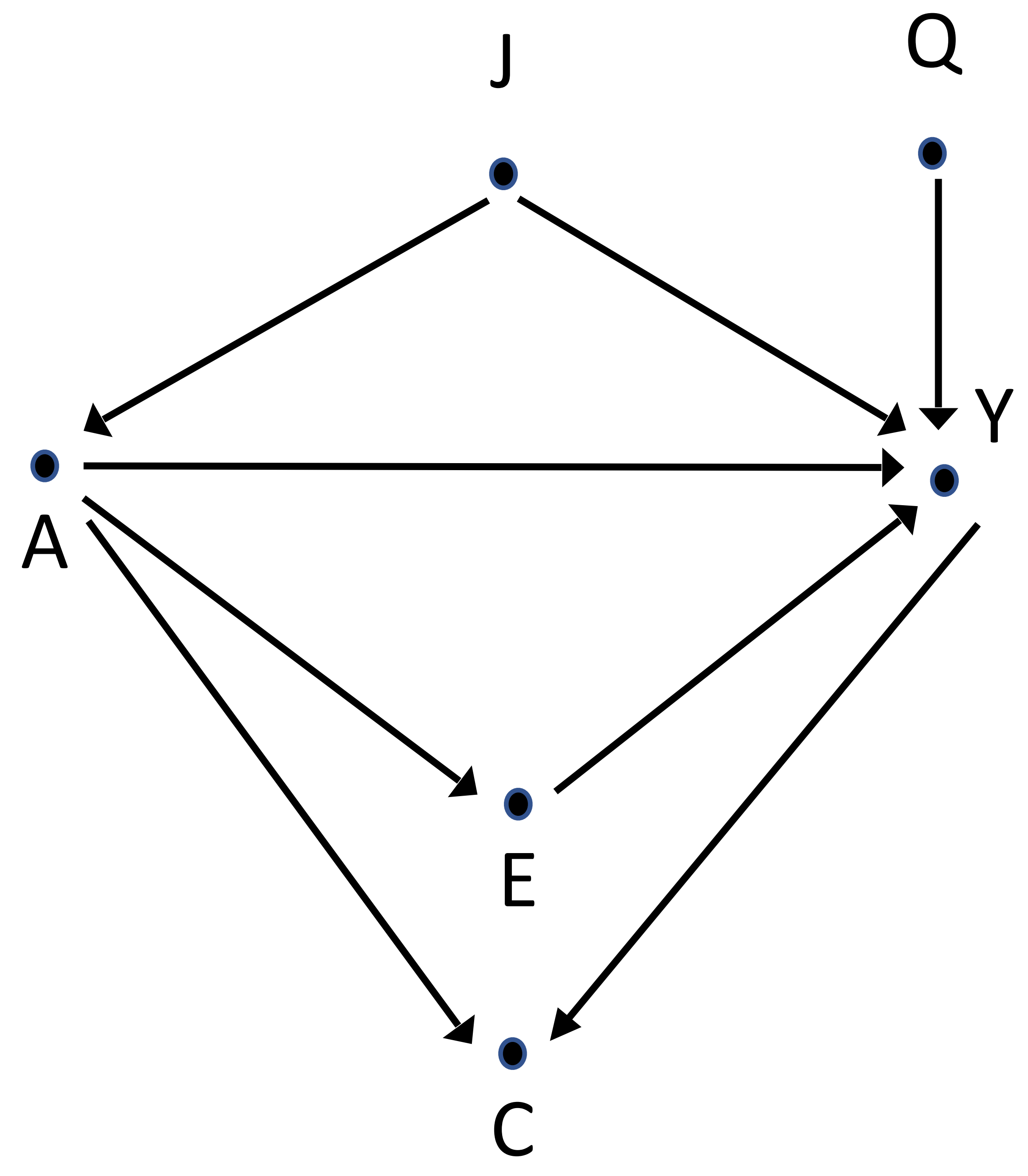
$P(y_{A=a})$

$P(y_{A\leftarrow a})$

$P(y_a)$

$P(y^a)$

How to measure the causal effect reliably ?



Causal Inference:

Estimating the effect of the intervention from observed data

$P(Y|do(A=a))$

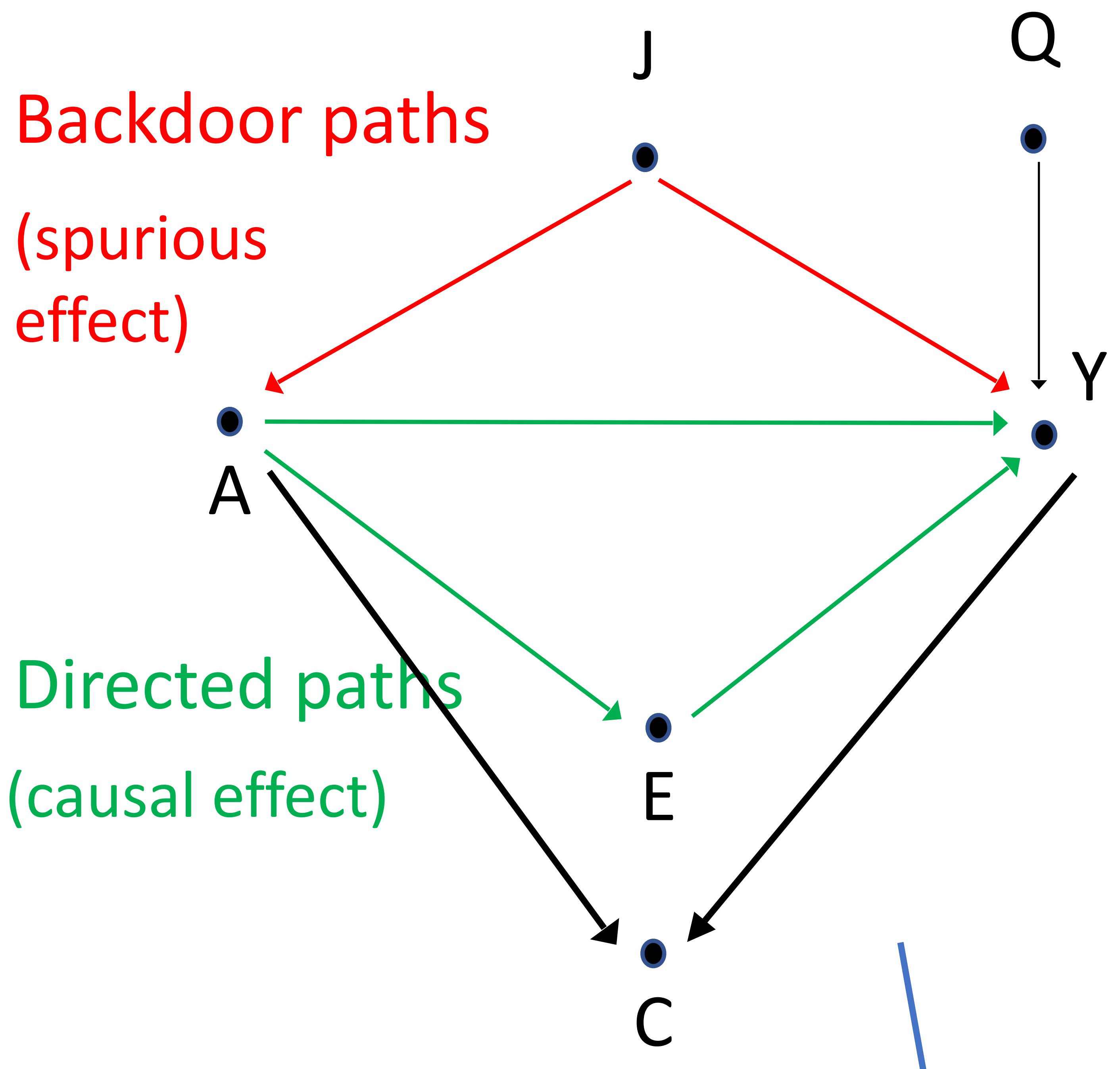
Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

* Pearl, J., *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009

How to measure the causal effect reliably ?



Causal Inference:

Estimating the effect of the intervention from observed data

P(Y|do(A=a))

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

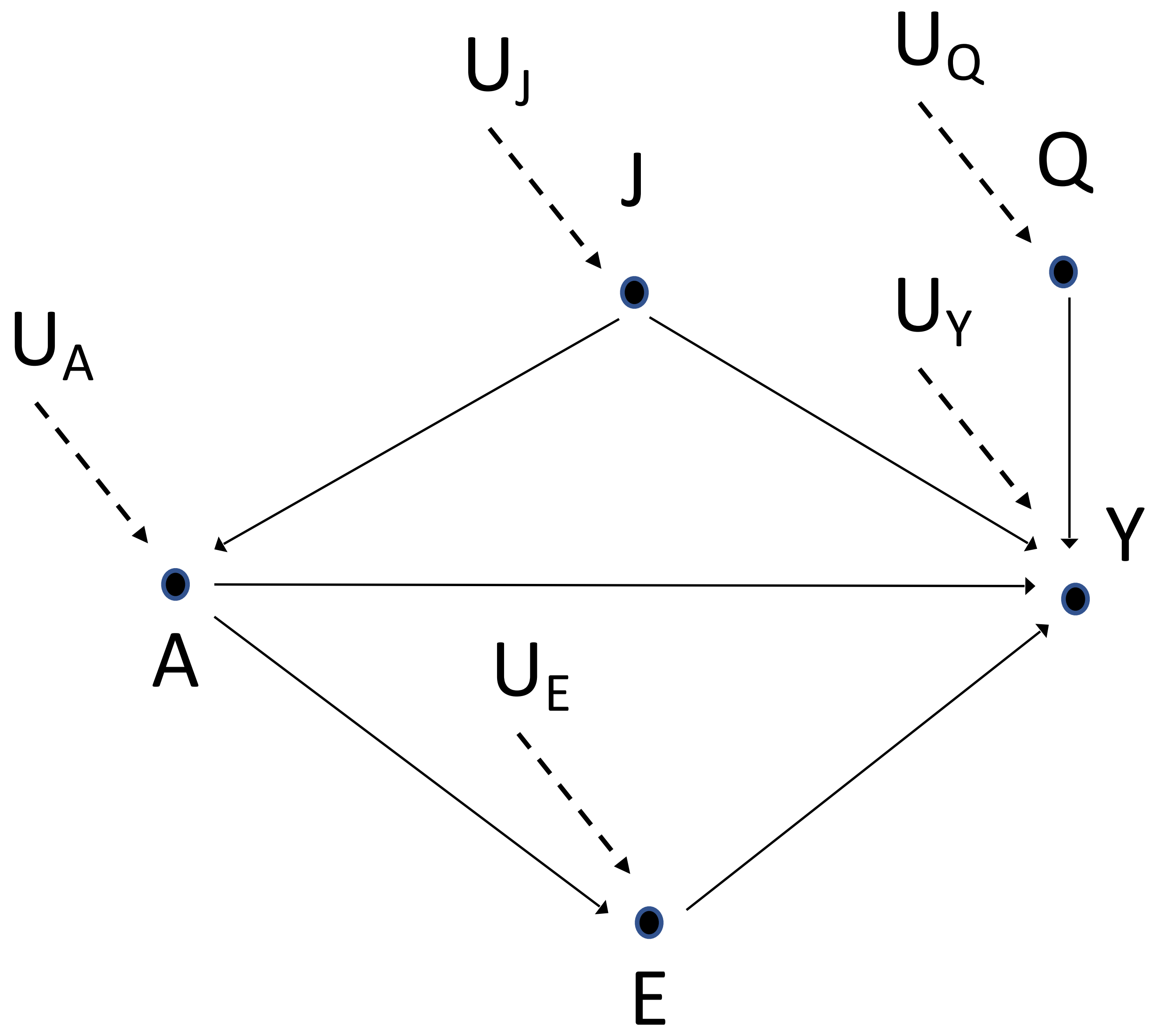
If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

$$P(Y = y|do(A = a)) = \sum_j P(Y = y|A = a, J = j)\, P(J = j)$$

* Pearl, J., *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009

How strong is the causal dependence of Y on A (causal effect of A on Y)?

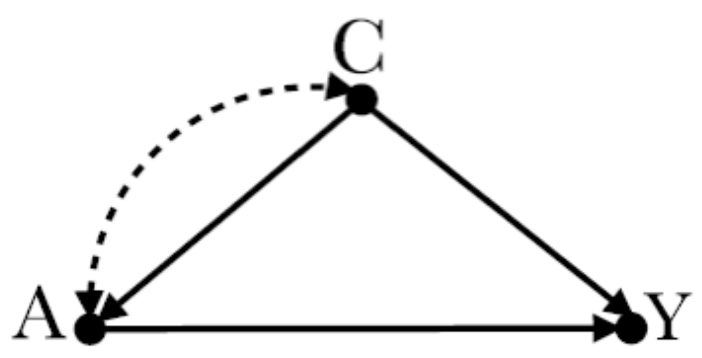


Markovian

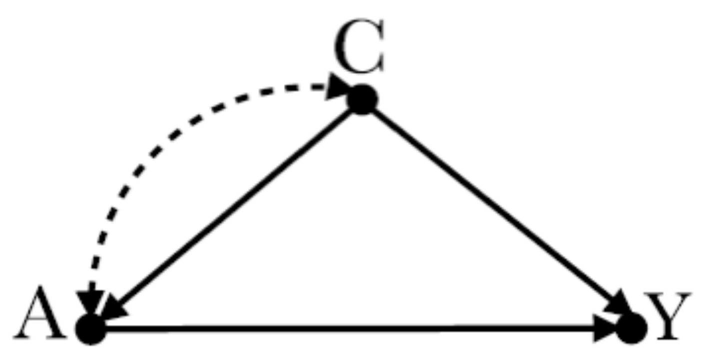
Estimating P(Y|do(A=a)) from observed data

Is it always possible ?

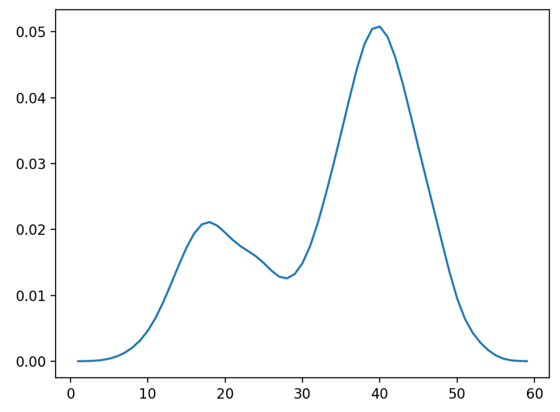
Identifiability



Causal model M1

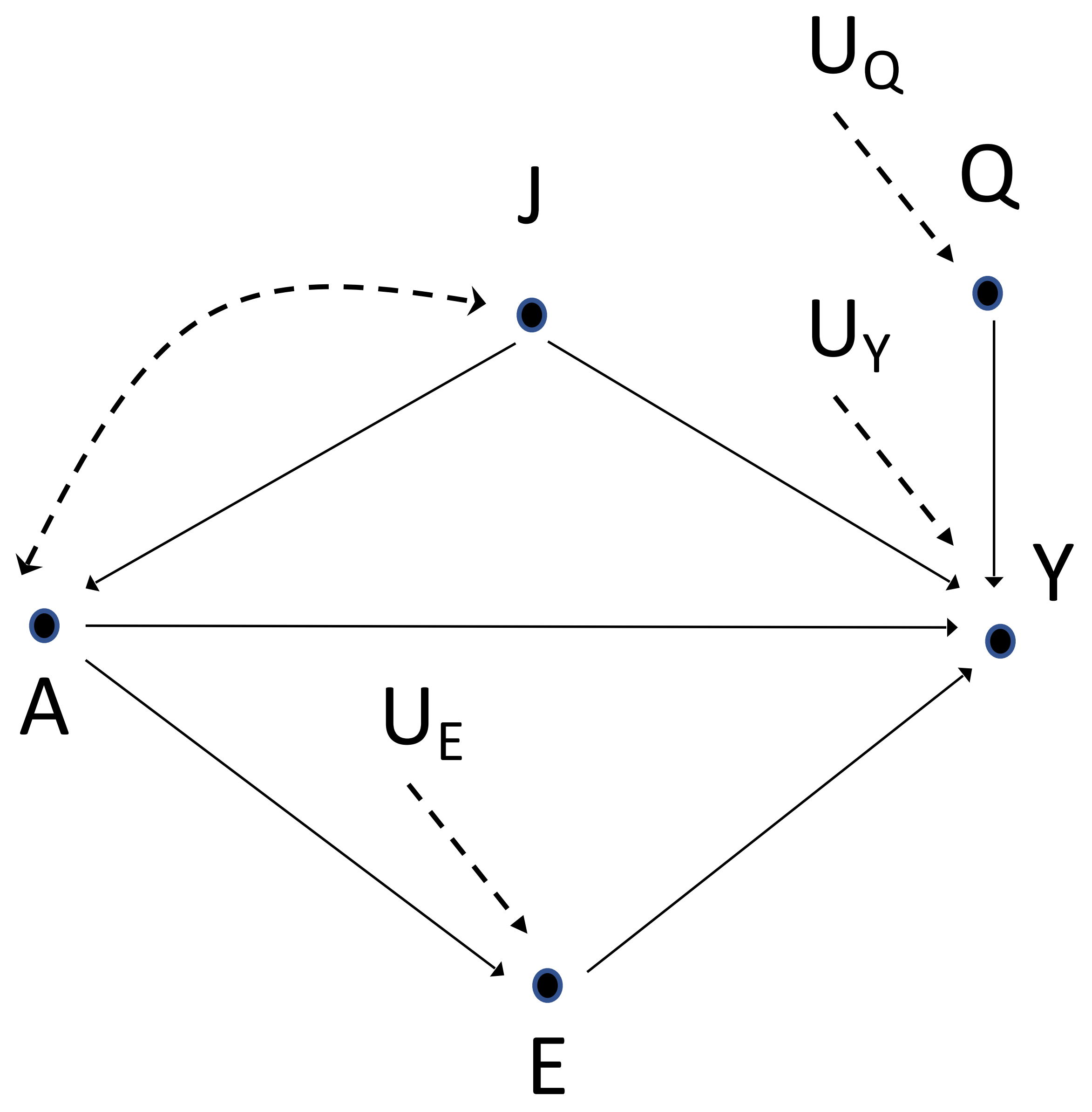


Causal model M2



Joint distribution

$P_{M1}(y|do(A=a))$ $\neq$ $P_{M2}(y|do(A=a))$



Semi-Markovian

How strong is the causal dependence of Y on A (causal effect of A on Y)?

Estimating P(Y|do(A=a)) from observed data in a semi-markovian model

Theorem 3.4.1 (Rules of *do* Calculus)
Let G be the directed acyclic graph associated with a causal model as defined in (3.2), and let P(·) stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W, we have the following rules.

Rule 1 (*Insertion/deletion of observations*):

$$P(y \mid \hat{x}, z, w) = P(y \mid \hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z) \mid X, W)_{G_{\overline{X}}}. \tag{3.31}$$

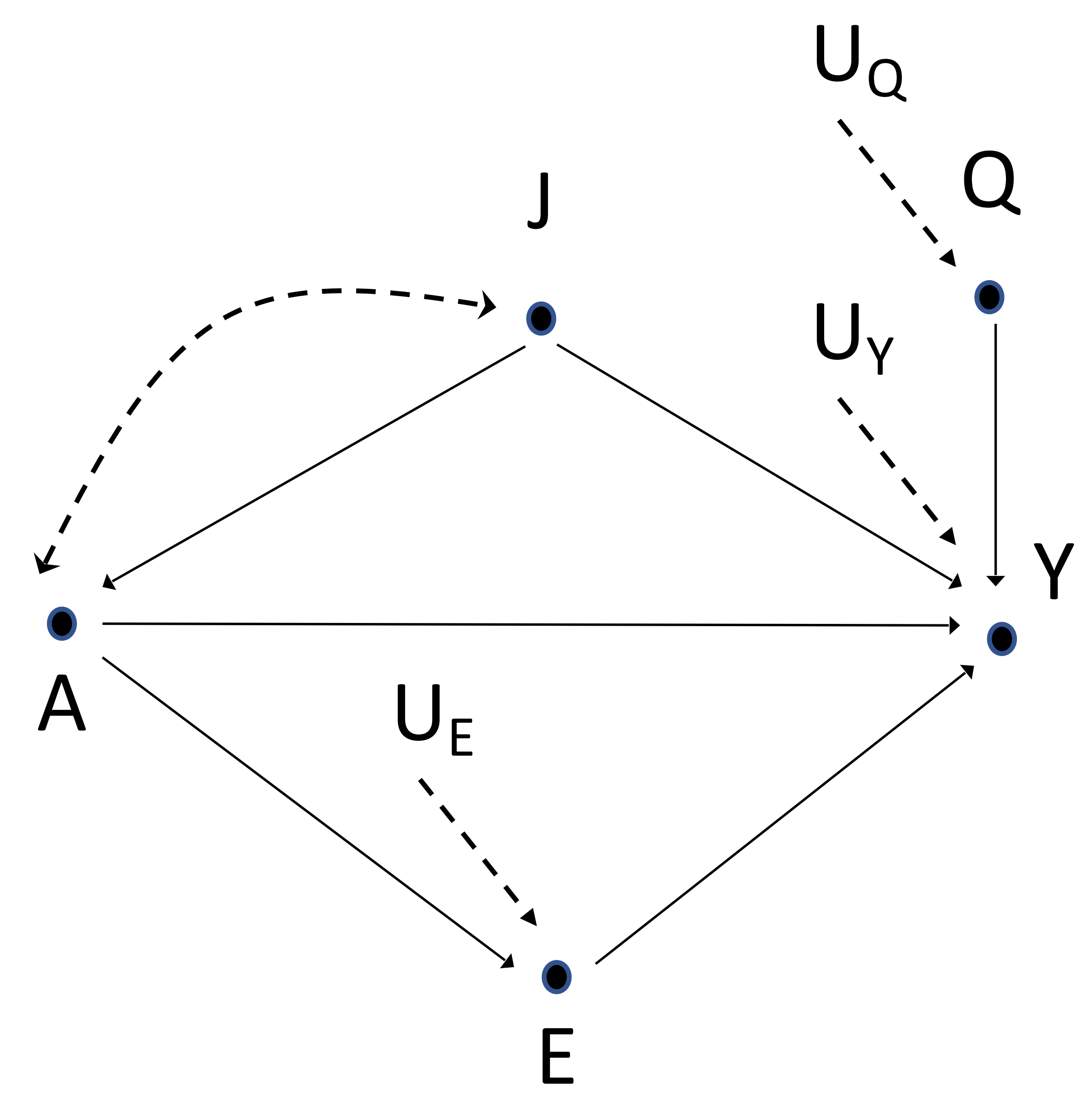
Rule 2 (*Action/observation exchange*):

$$P(y \mid \hat{x}, \hat{z}, w) = P(y \mid \hat{x}, z, w) \quad \text{if } (Y \perp\!\!\!\perp Z) \mid X, W)_{G_{\overline{X}\underline{Z}}}. \tag{3.32}$$

Rule 3 (*Insertion/deletion of actions*):

$$P(y \mid \hat{x}, \hat{z}, w) = P(y \mid \hat{x}, w) \text{ if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}, \overline{Z(W)}}}, \tag{3.33}$$

where Z(W) is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

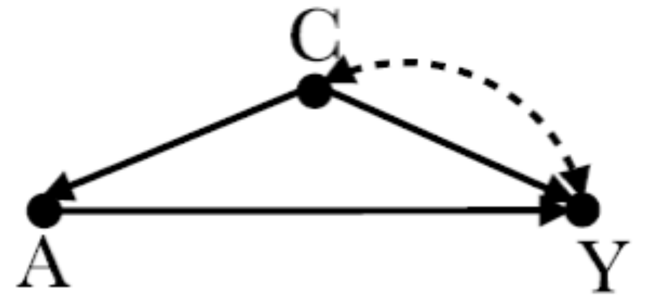


Semi-Markovian

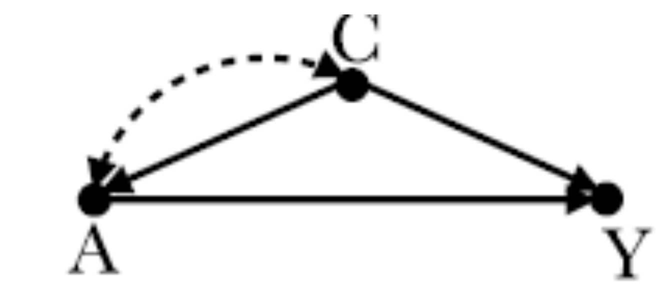
How strong is the causal dependence of Y on A (causal effect of A on Y)?

Estimating P(Y|do(A=a)) from observed data in a semi-markovian model

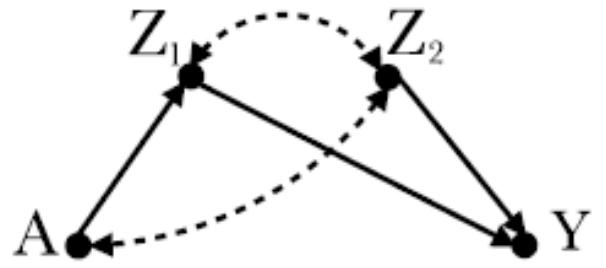
Graphical criterion: If the cause variable (X or A) is not connected to any of its direct children through a confounding path, it is identifiable.

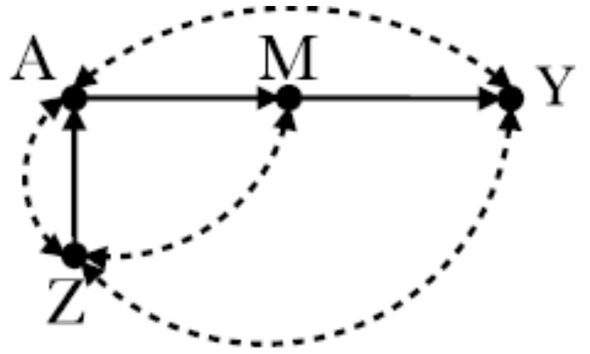


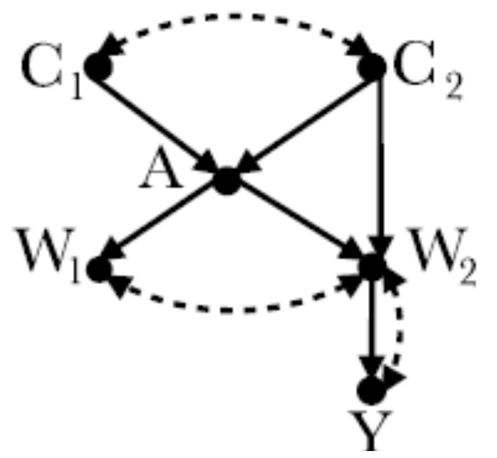
$$\sum_C P(y|a,c)\ P(c)$$



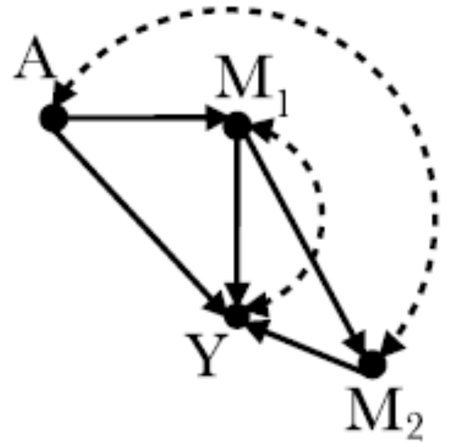
$$\sum_C P(y|a,c)\ P(c)$$





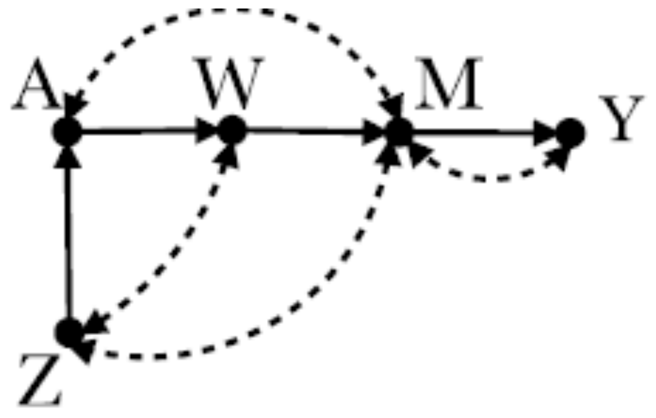


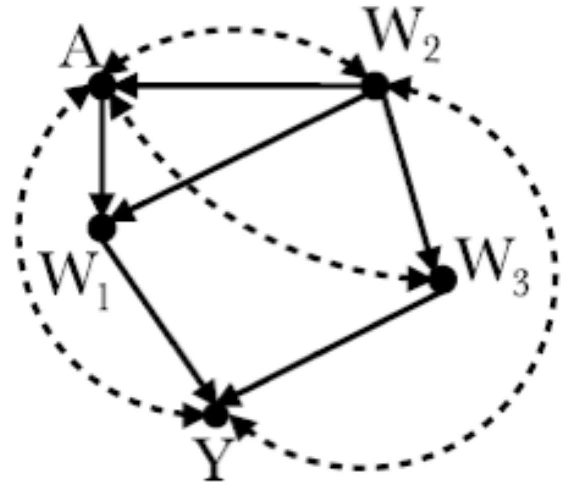
$$\sum_{c_1,c_2} P(y|a,c_1,c_2)\ P(c_1,c_2)$$



$$\sum_{m_1,m_2}\ P(y|m_1,m_2,a)\ P(m_1|a)$$

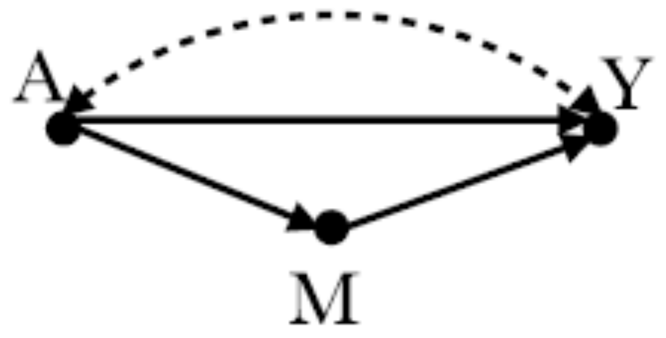
$$\times \sum_{a'}\ P(m_2|m_1,a')\ P(a')$$





$$\sum_{w_1}\sum_{w_2}\sum_{a'}\ P(y|w_1,w_2,a')\ P(a'|w_2)$$

$$\times P(w_1|w_2,a)P(w_2)$$



$$\sum_M P(y|m,a)\ P(a)\ P(m|a)$$

Front-door criterion

Survey papers about Fairness and Causality

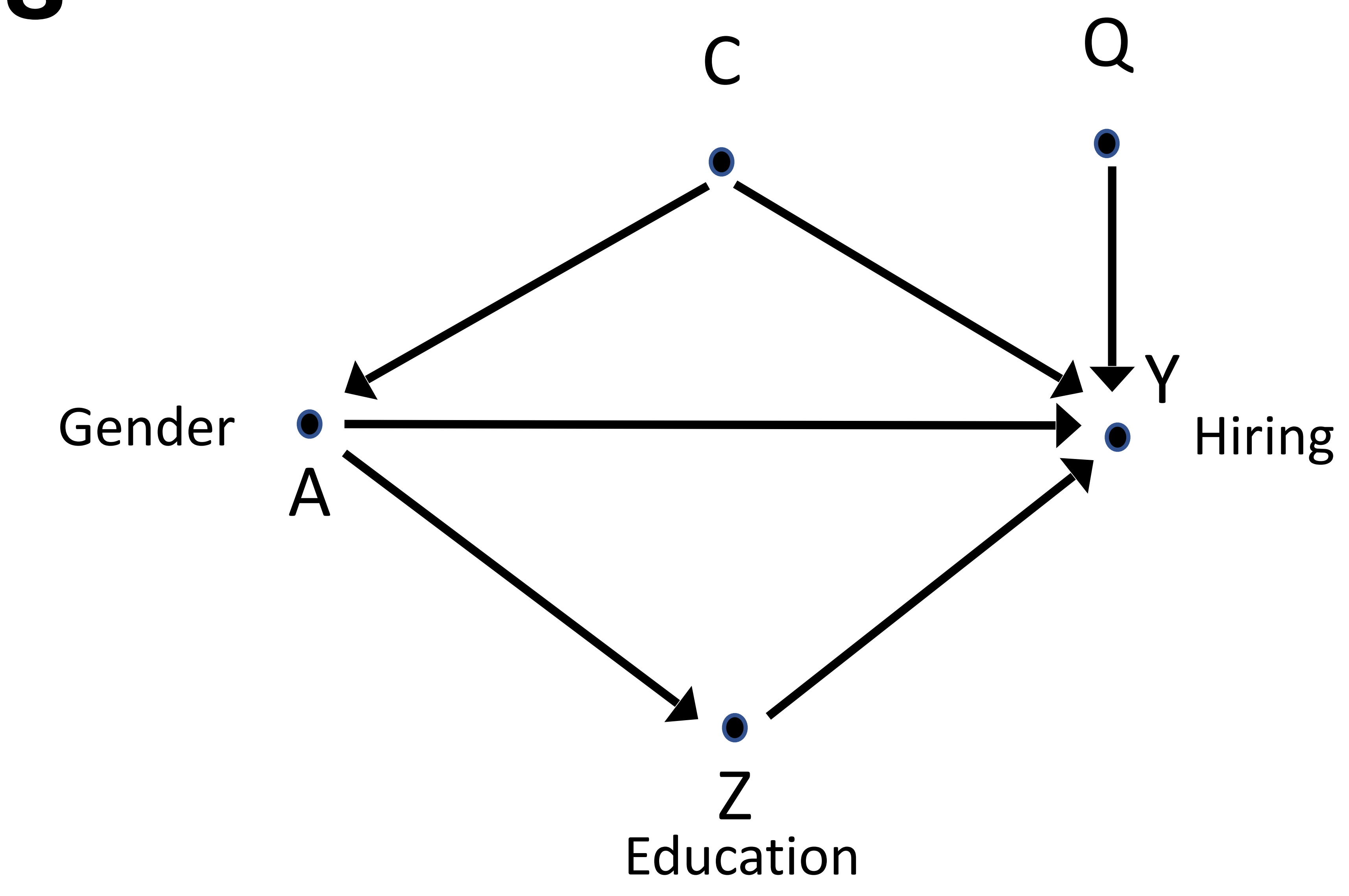
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021).
Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, *58*(5), 102642.

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2022).
Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*. (Under review)

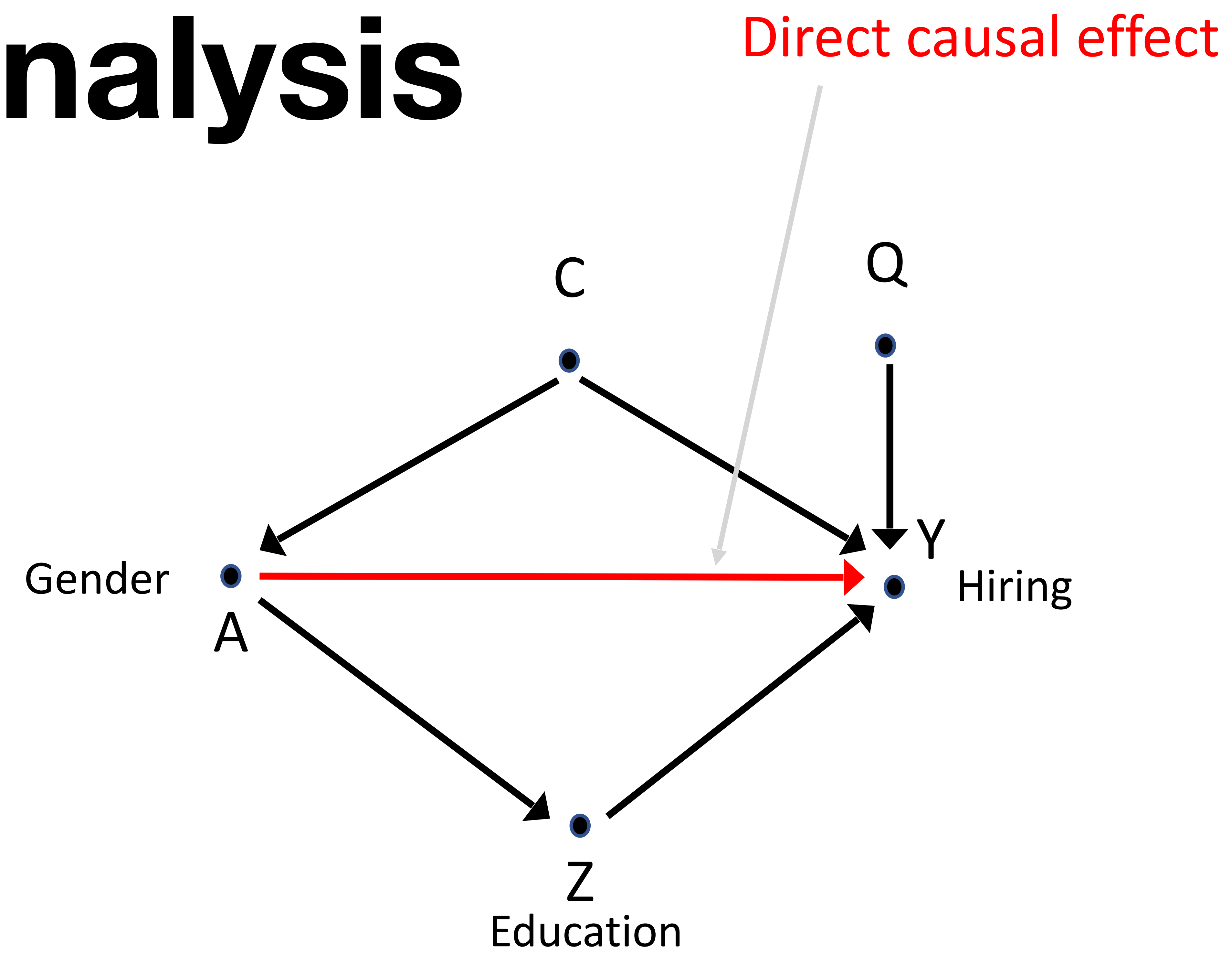
Makhlouf, K., Zhioua, S., & Palamidessi, C. (2022, December).
Identifiability of Causal-based ML Fairness Notions. In *2022 14th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 1-8). IEEE.

Causality Benefit 2:

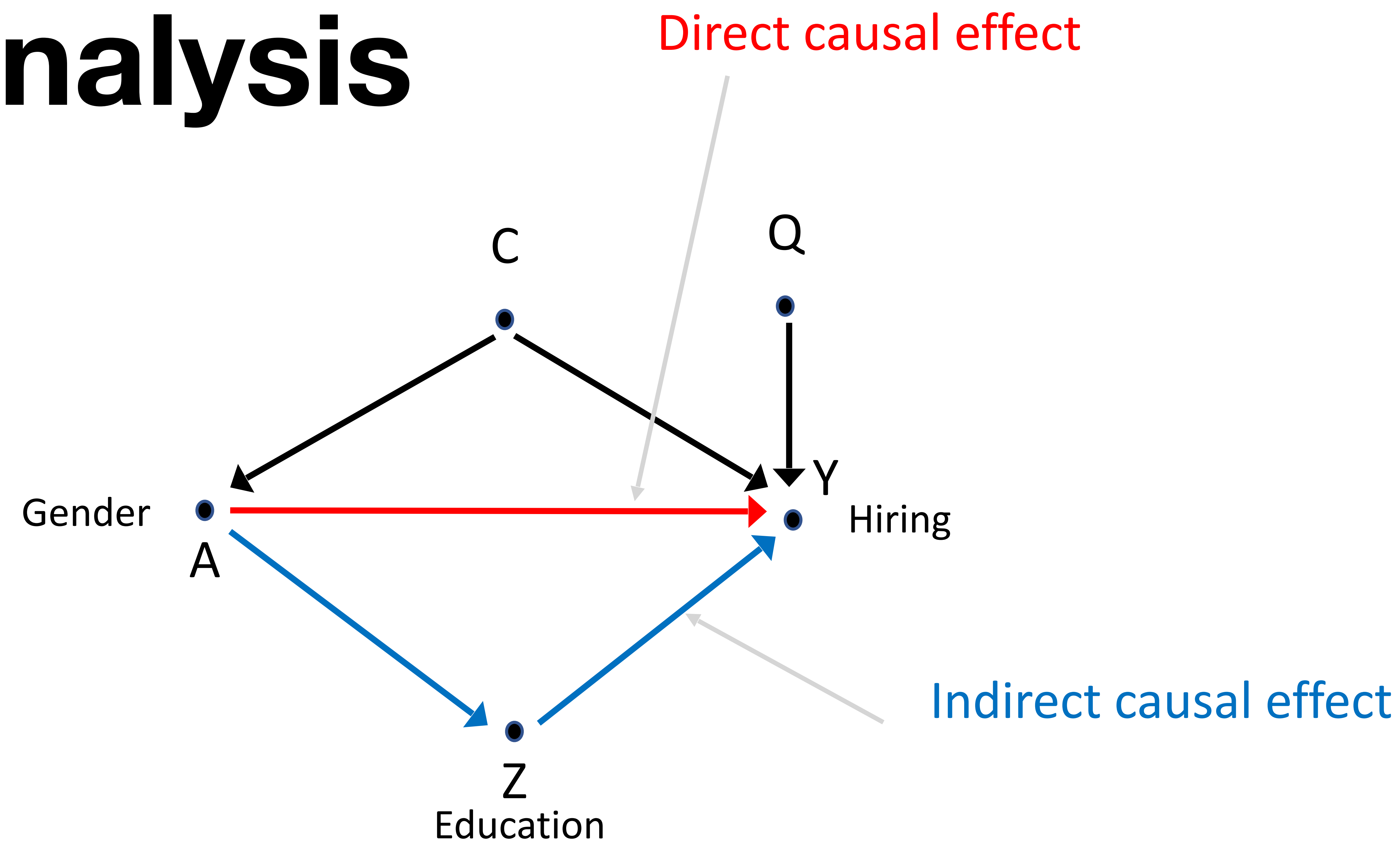
Mediation Analysis



Mediation Analysis



Mediation Analysis

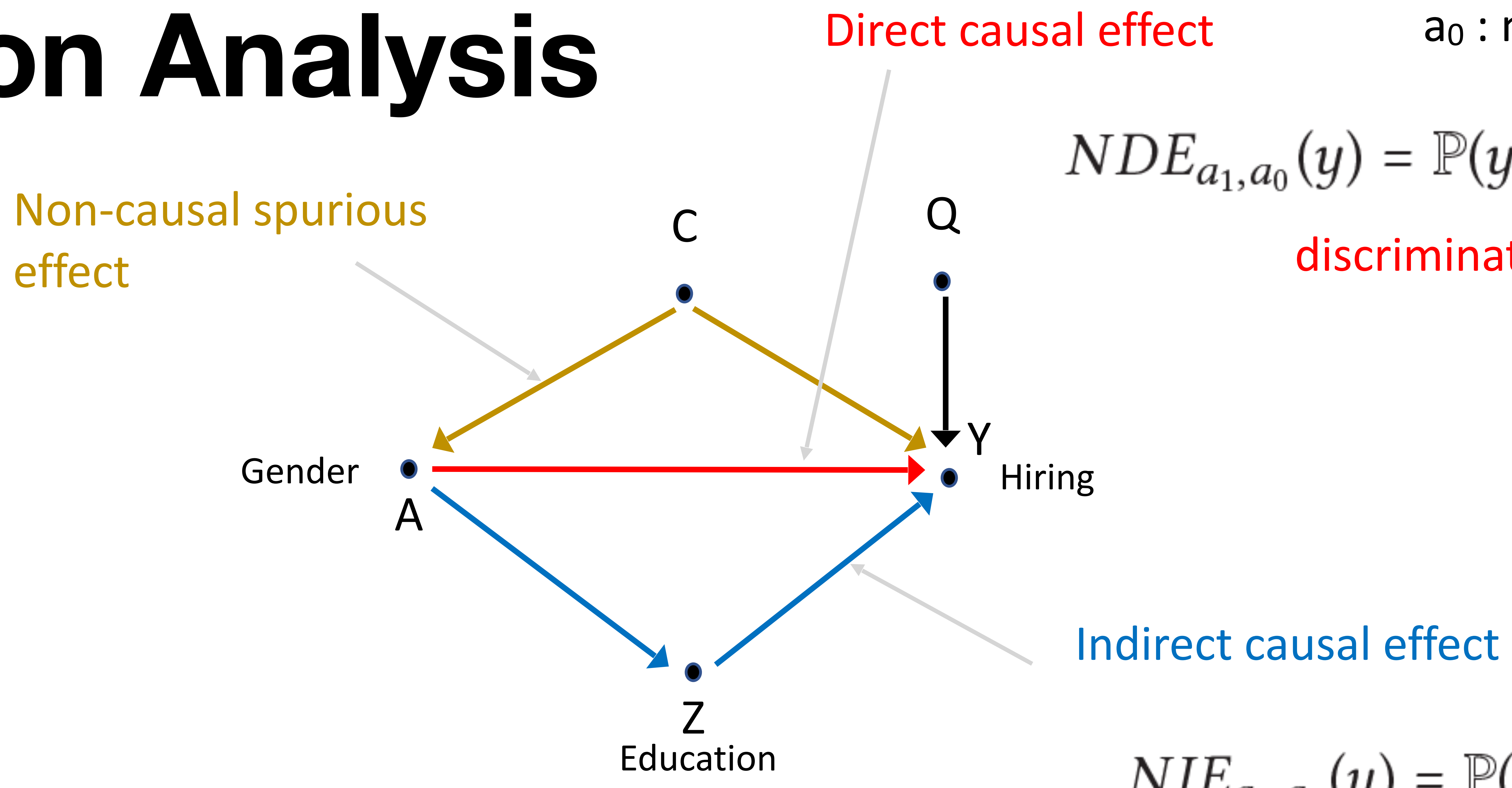


$P(y_a) = P(Y=y|do(A=a))$

a_1 : female

a_0 : male

Mediation Analysis



$$NDE_{a_1,a_0}(y) = \mathbb{P}(y_{a_1,Z_{a_0}}) - \mathbb{P}(y_{a_0})$$

discrimination

$$NIE_{a_1,a_0}(y) = \mathbb{P}(y_{a_0,Z_{a_1}}) - \mathbb{P}(y_{a_0})$$

Discrimination ? It depends on Z

* Pearl, J. (2001). Direct and indirect effects. In Proceeding of UAI 2001.

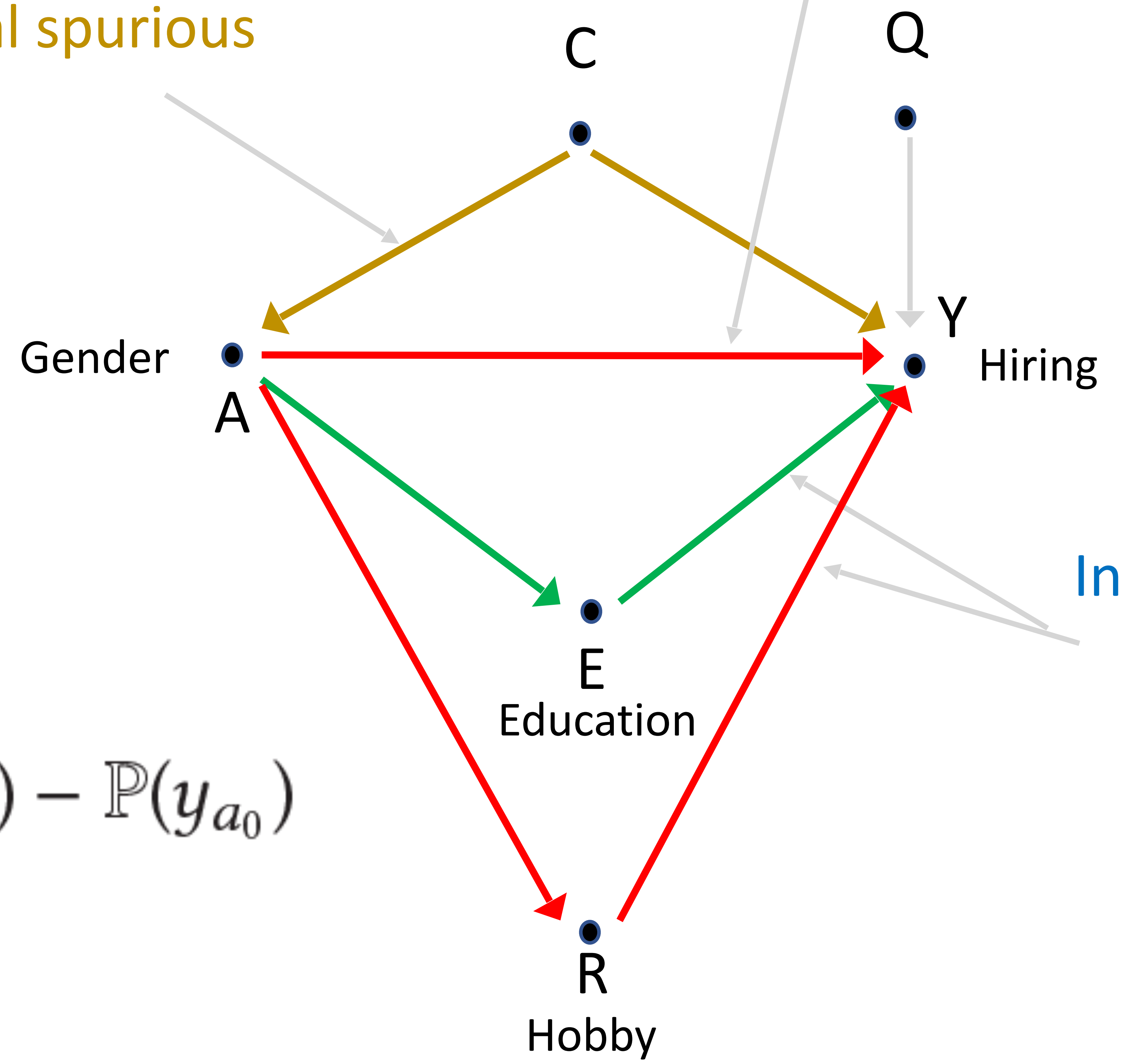
Mediation Analysis

Direct causal effect

$$NDE_{a_1,a_0}(y) = \mathbb{P}(y_{a_1,Z_{a_0}}) - \mathbb{P}(y_{a_0})$$

discrimination

Non-causal spurious effect



Indirect causal effect

$$NIE_{a_1,a_0}(y) = \mathbb{P}(y_{a_0,Z_{a_1}}) - \mathbb{P}(y_{a_0})$$

Discrimination ? It depends on Z

Path-Specific effect

$$PSE^{\pi}_{a_1,a_0}(y) = \mathbb{P}(y_{a_1|_{\pi},a_0|_{\overline{\pi}}}) - \mathbb{P}(y_{a_0})$$

* Pearl, J. (2001). Direct and indirect effects. In Proceeding of UAI 2001.

* Chiappa, S. (2019). Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 7801-7808).

Simpson's Paradox

Statistical parity = 7/15 – 8/15 = **-1/15**

Discrimination against women

Discrimation in favor of women

Hiring rate (T = 0) 3/10 = 0.3

Hiring rate (T = 1) 4/5 = 0.8

Total hiring rate 7/15

	A	T	Ŷ
	Gender	Job Type	Hiring
A=1 (Women)	1	0	1
	1	0	1
	1	0	1
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	0	0
	1	1	1
	1	1	1
	1	1	1
	1	1	1
	1	1	0

	A	T	Ŷ
	Gender	Job Type	Hiring
A=0 (Men)	0	0	1
	0	0	0
	0	0	0
	0	0	0
	0	0	0
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	1
	0	1	0
	0	1	0
	0	1	0

Hiring rate (T=0) 1/5 = 0.2

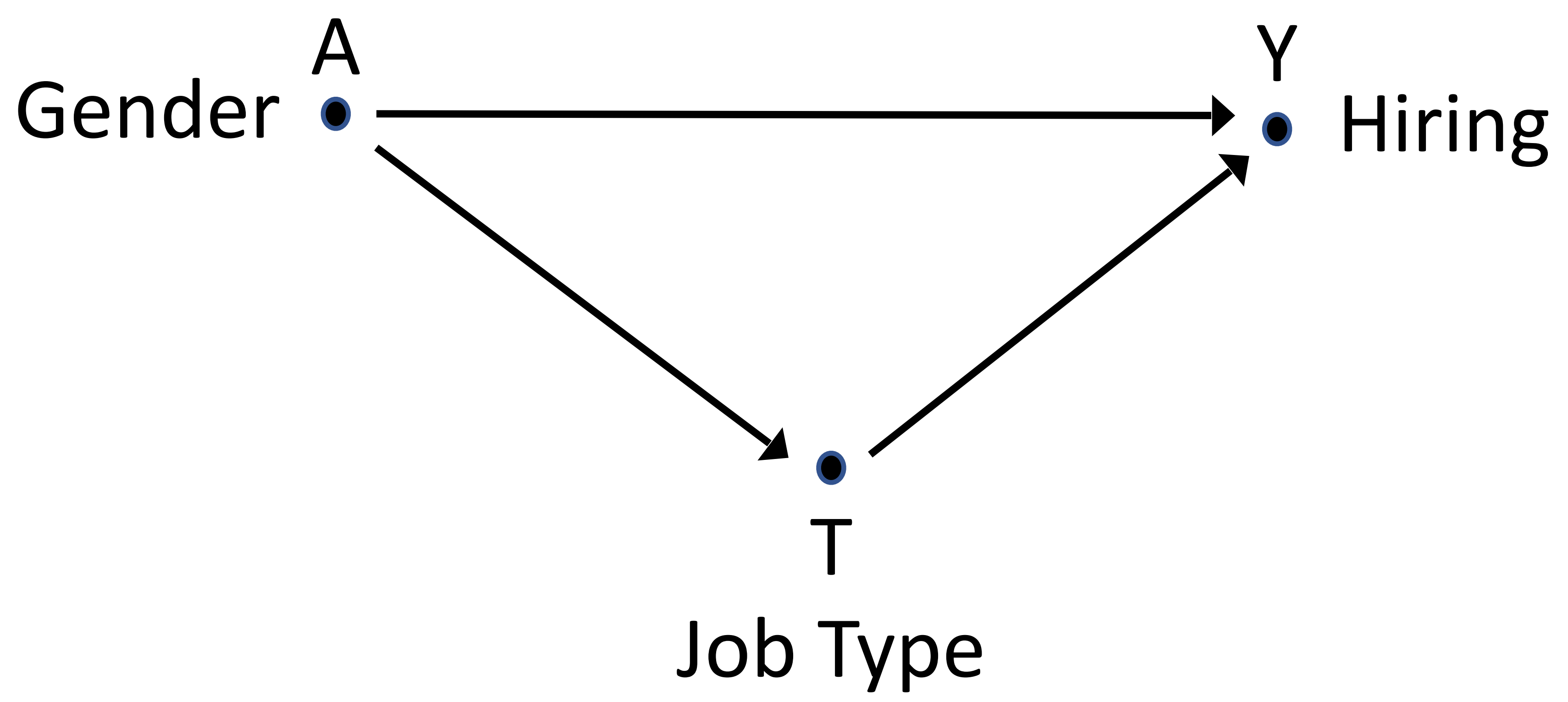
Hiring rate (T=1) 7/10 = 0.7

Total hiring rate 8/15

A = 0	Man
A = 1	Woman

T = 0	Flexible time job
T = 1	Non-flexible time job

Y=0	Not hired
Y=1	Hired


A
Gender
Y
Hiring
T
Job Type

Dissecting Bias

- **Bias**: “deviation of the expected value from the quantity it estimates”

Example: $\mathbb{E}[\hat{Y}_{\mathcal{S}}] - \mathbb{E}[Y]$

- **Discrimination**: “unjust or prejudicial treatment of different categories of people, on the ground of race, age, gender, disability, religion, political belief, etc.

Example: $\mathbb{E}[Y|A = a_1] - \mathbb{E}[Y|A = a_0]$

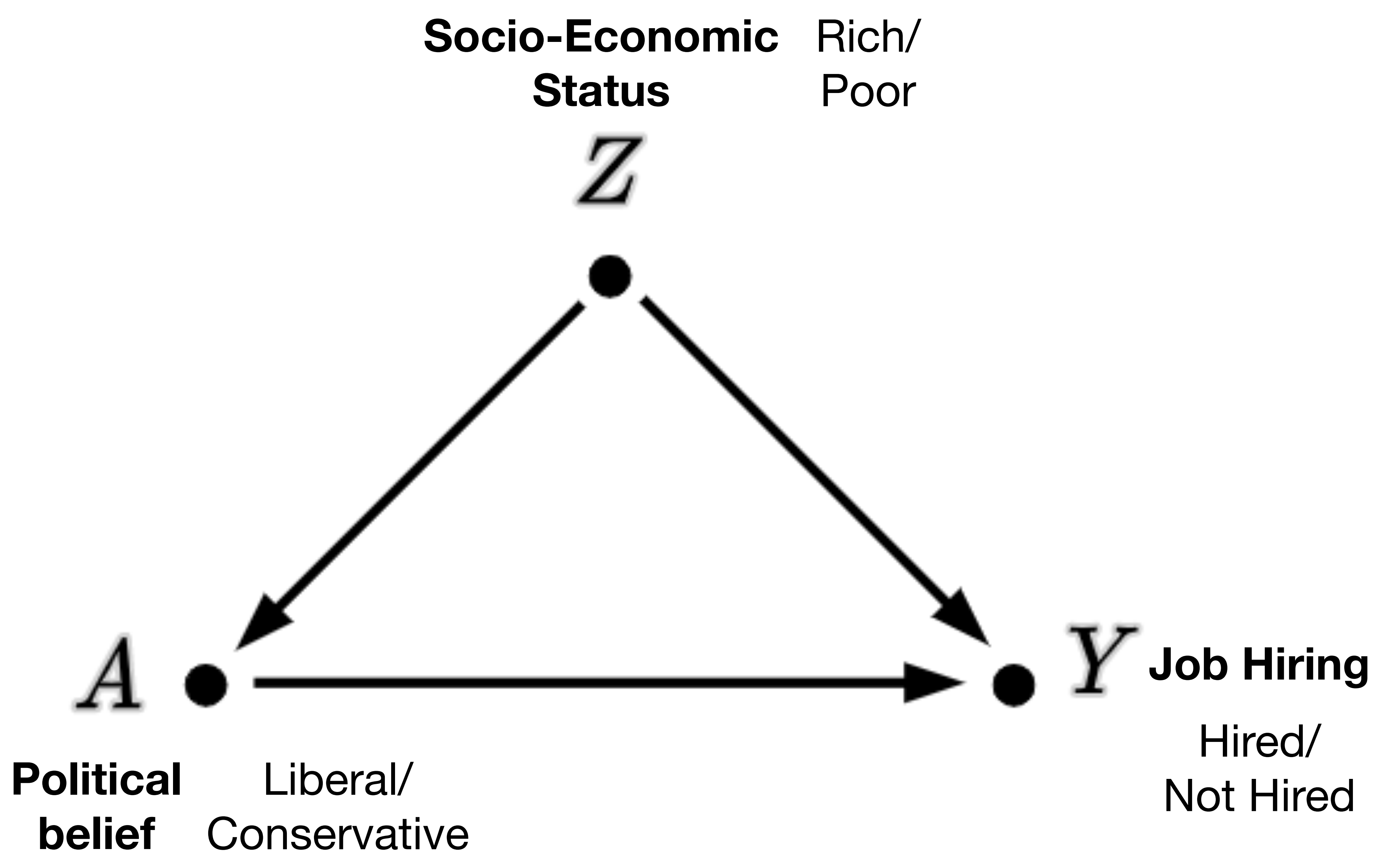
- A bias in measuring discrimination may **amplify** or **under-estimate** the true discrimination

Dissecting Bias

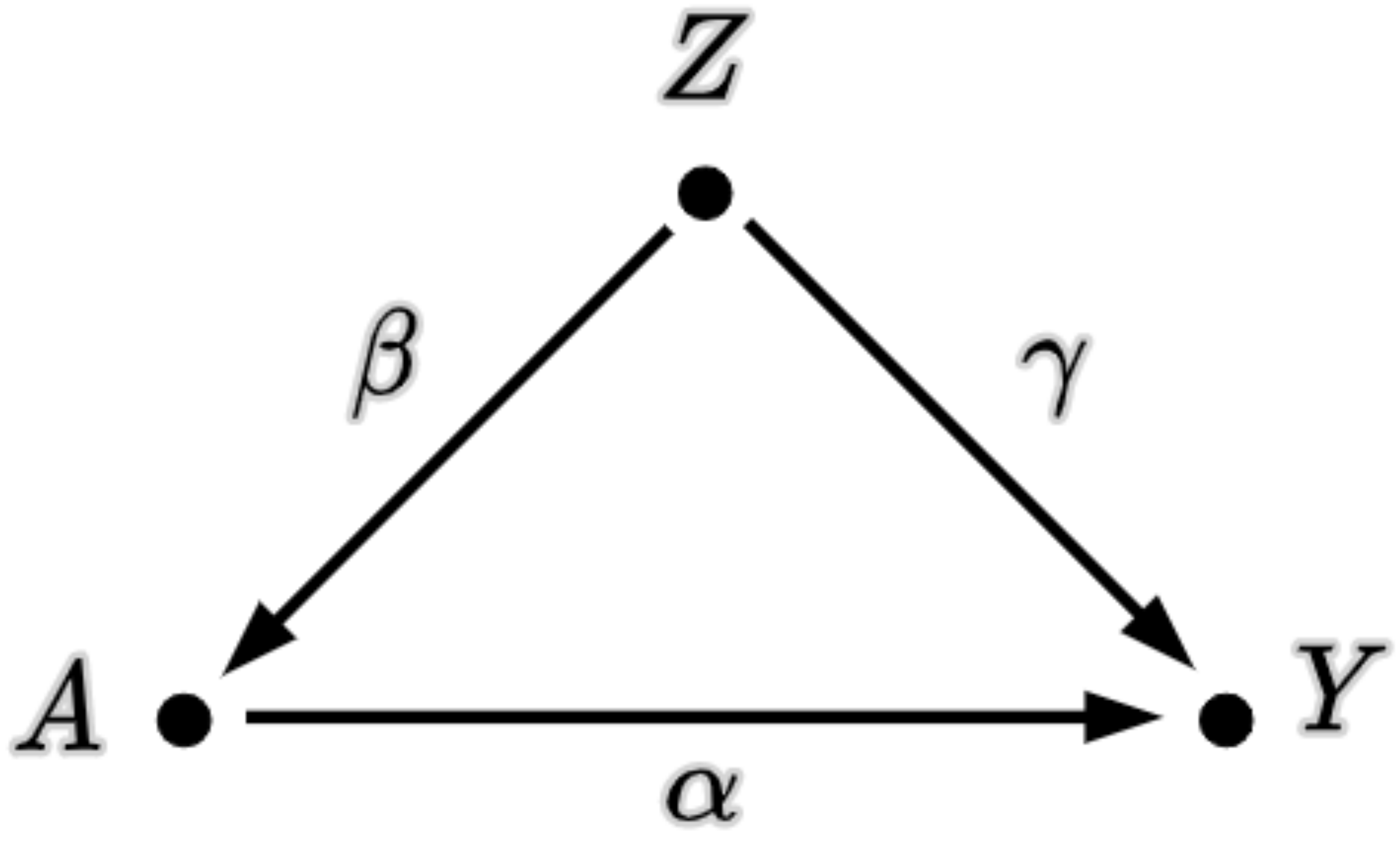
- ***Confounding Bias***: failing to identify and adjust on a confounder
- ***Collider (Selection) Bias***: implicit adjustment on a collider
- ***Measurement Bias***: adjusting on a proxy variable
- ***Representation Bias***: due to under-representation of sub-populations

Confounding Bias

Failing to adjust on confounder(s)



Confounding Bias (Linear case)

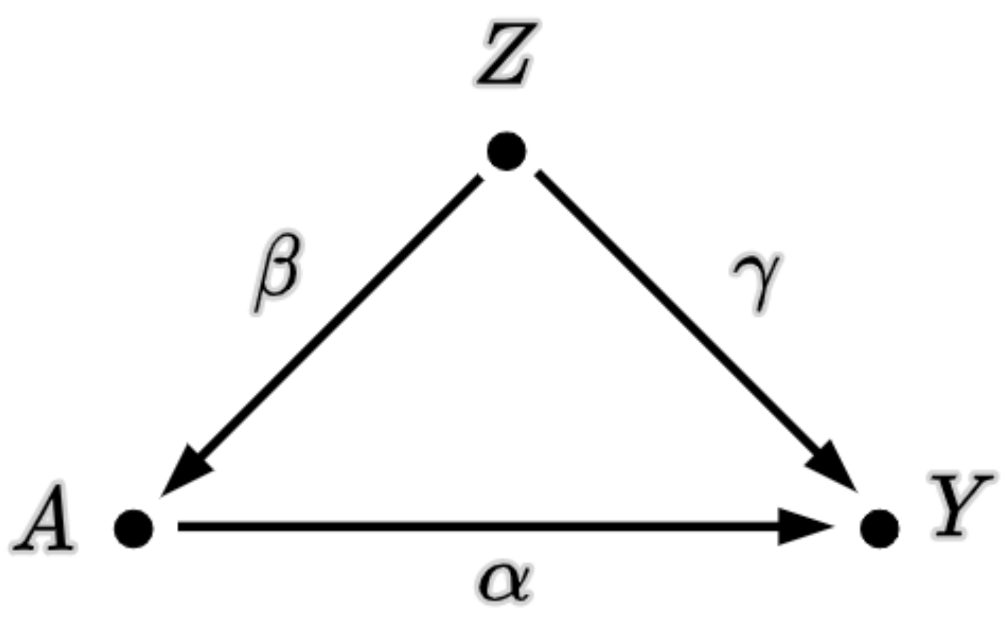


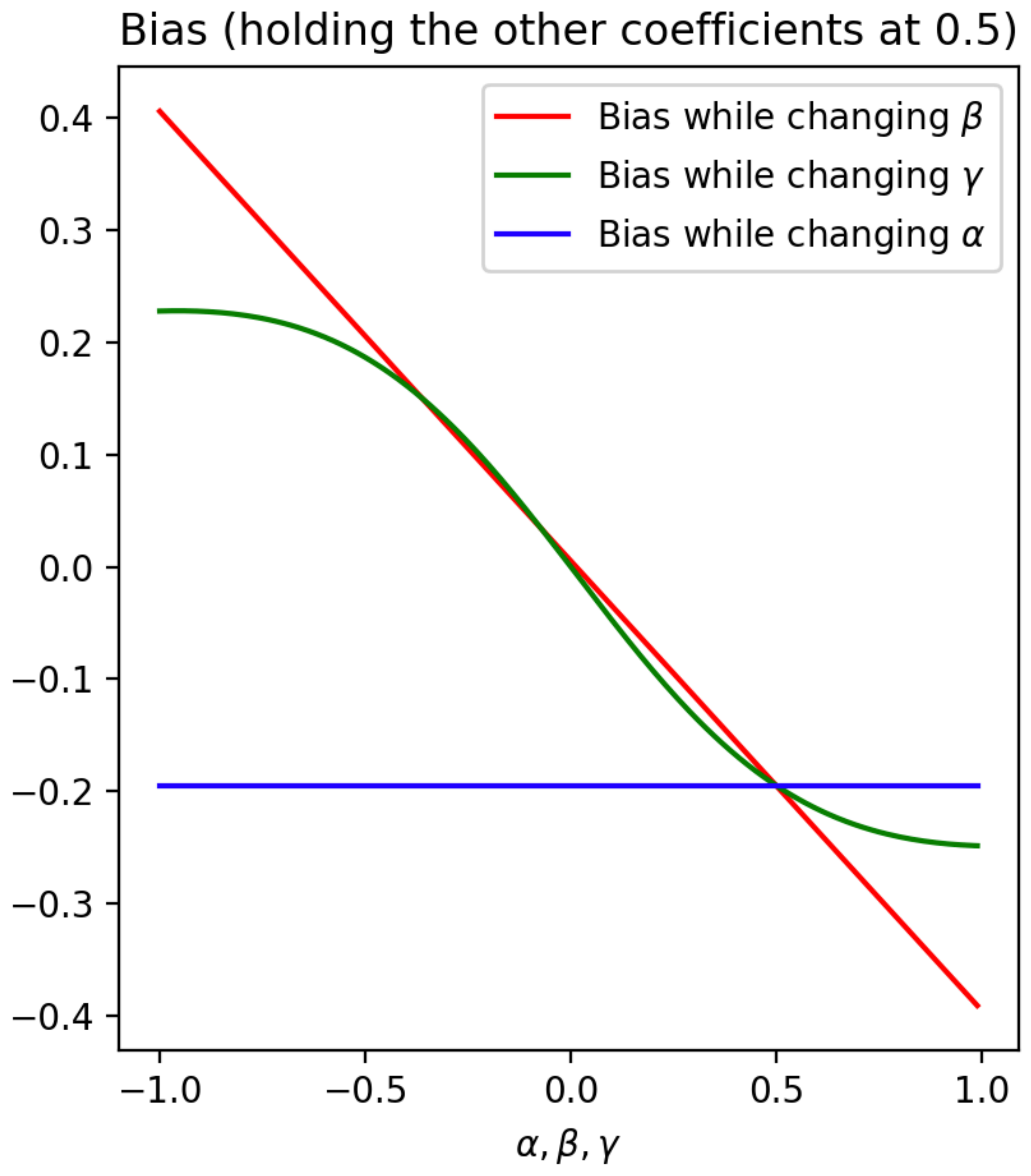
$$
\begin{aligned}
ConfBias(Y, A) &= \beta_{ya} - \beta_{ya.z} \\
&= \frac{\sigma_{az}\left(\sigma_{yz} - \frac{\sigma_{ya}}{{\sigma_a}^2}\sigma_{az}\right)}{{\sigma_a}^2{\sigma_z}^2 - {\sigma_{az}}^2} \\
&= \frac{{\sigma_z}^2}{{\sigma_a}^2}\beta\gamma
\end{aligned}
$$

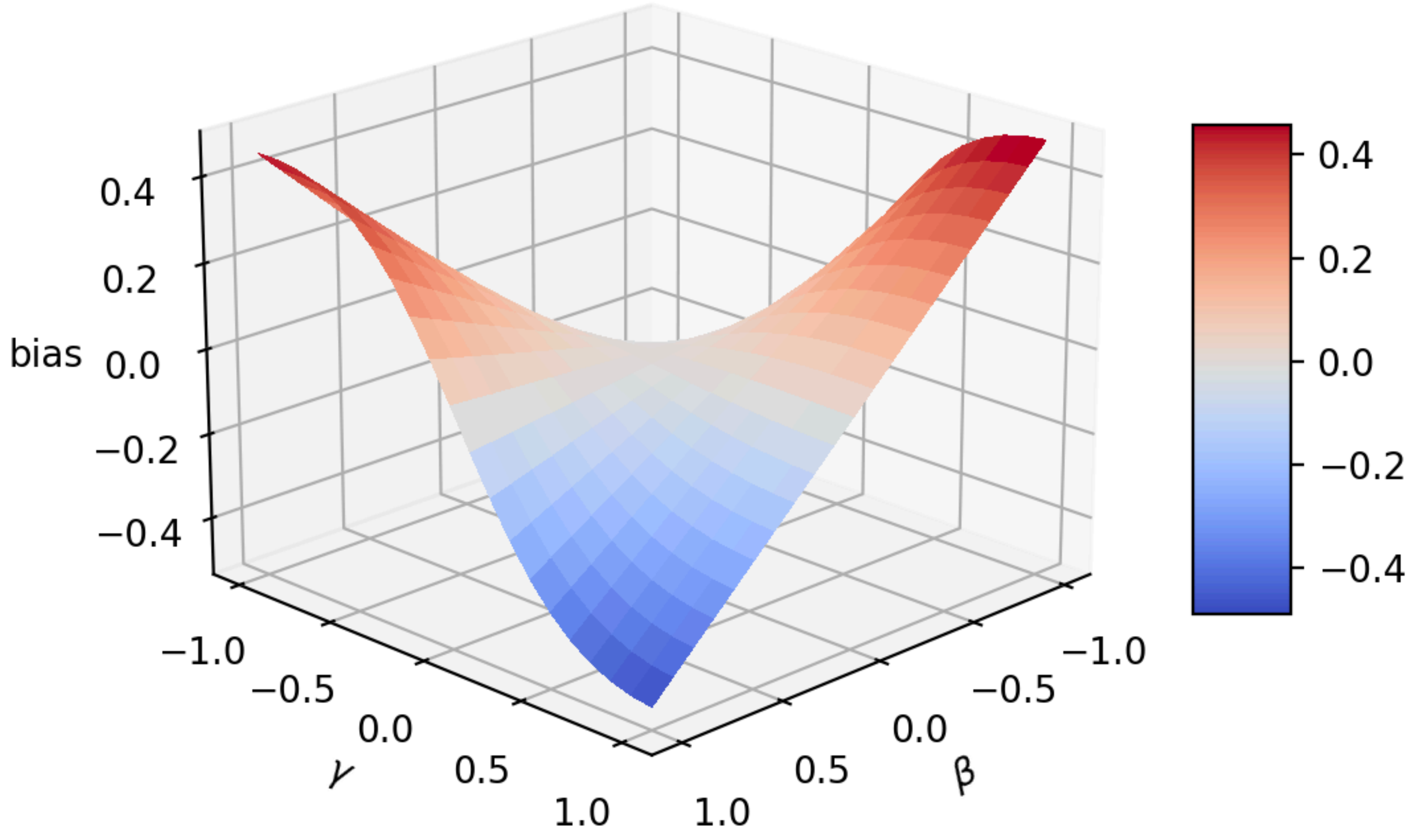
Proof using results from:

Cramér, H. (1999). *Mathematical methods of statistics* (Vol. 26). Princeton university press.
Wright, S. Correlation and causation. Journal of Agricultural Research, 20:557–585, 1921

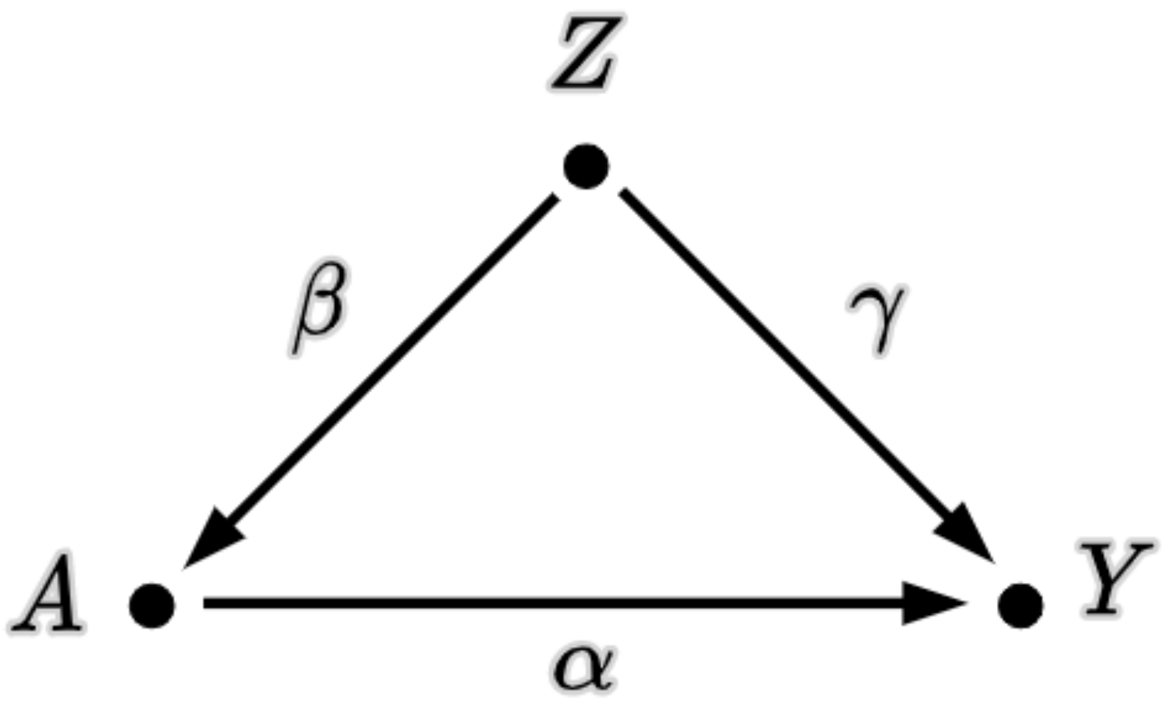
Confounding Bias (Linear case)

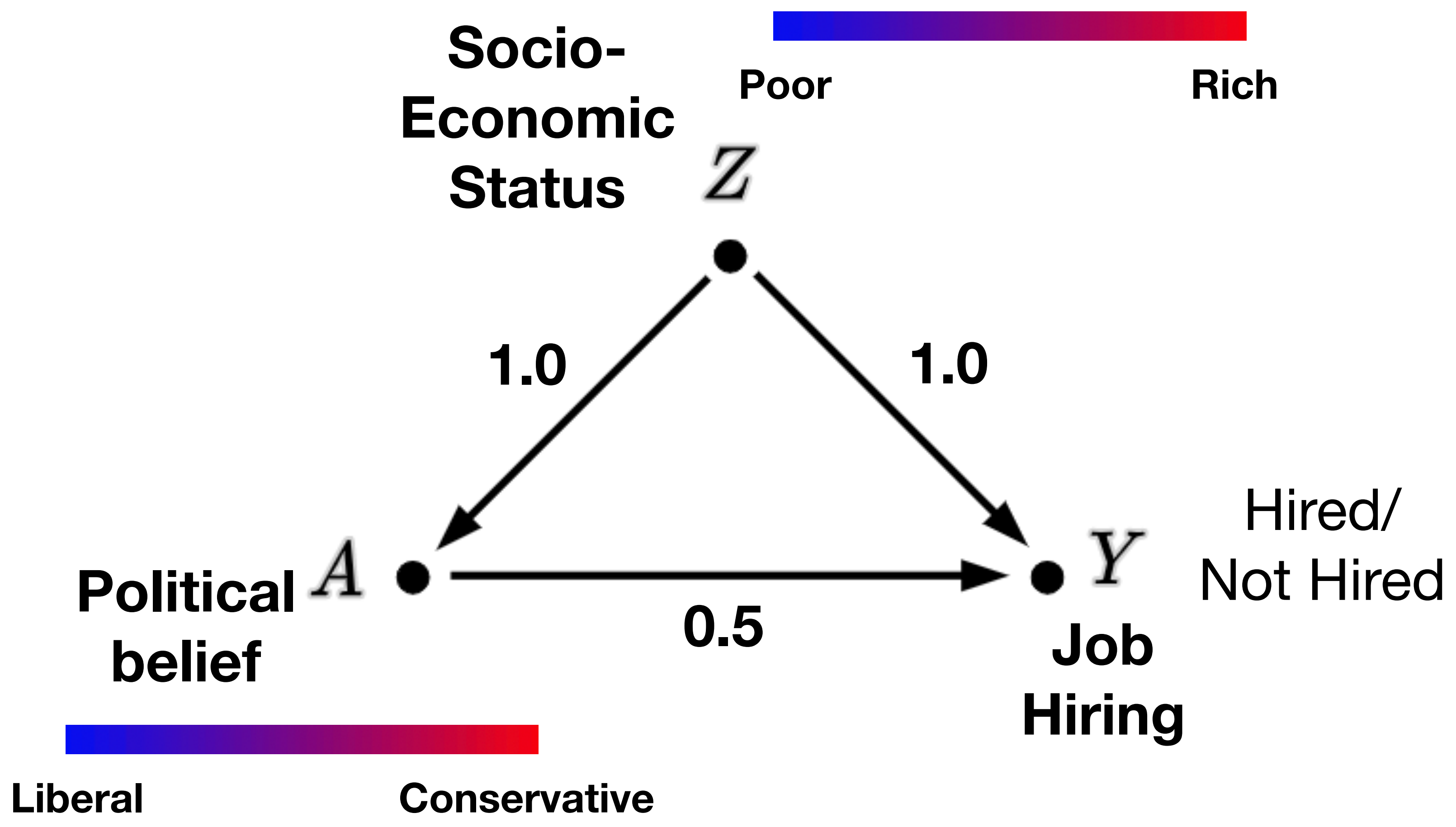


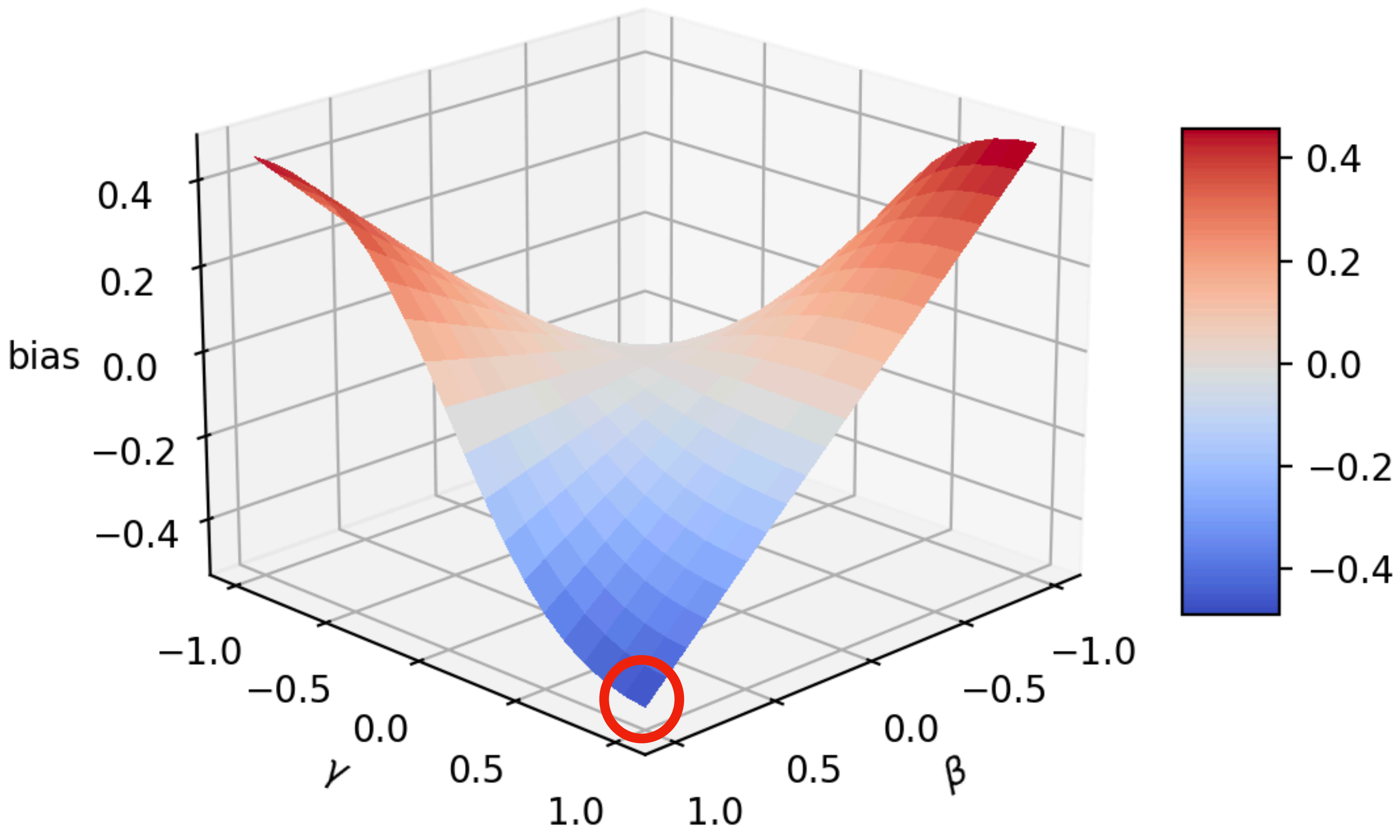




Confounding Bias (Linear)

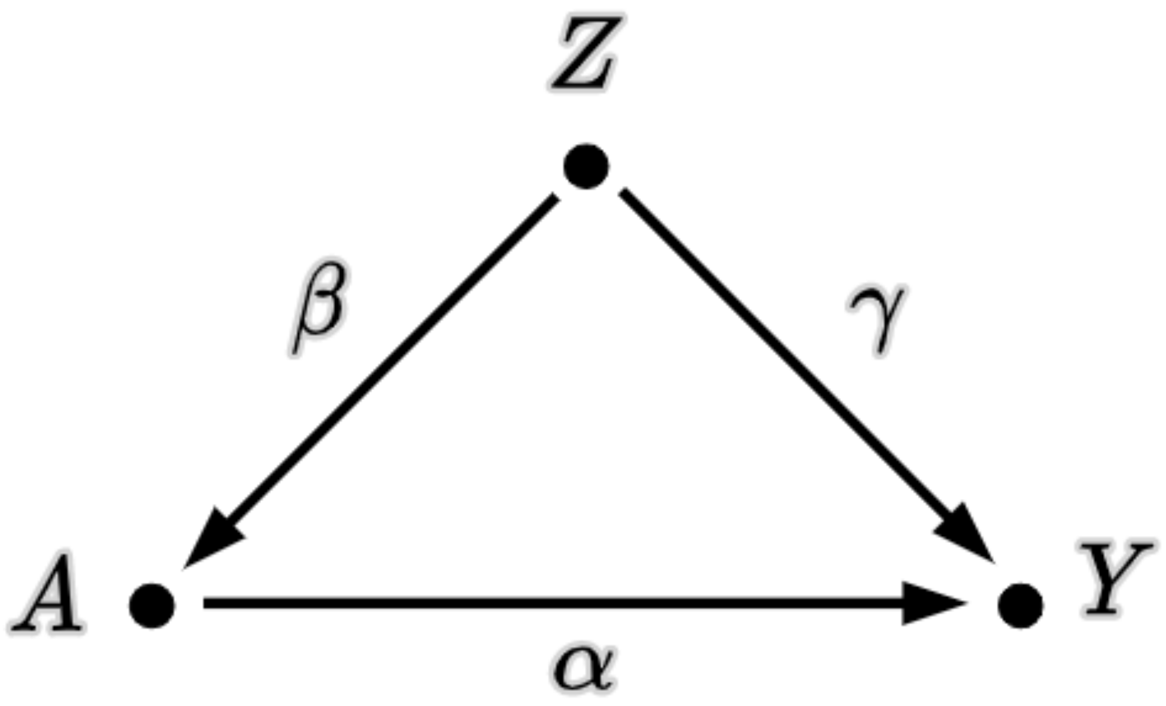


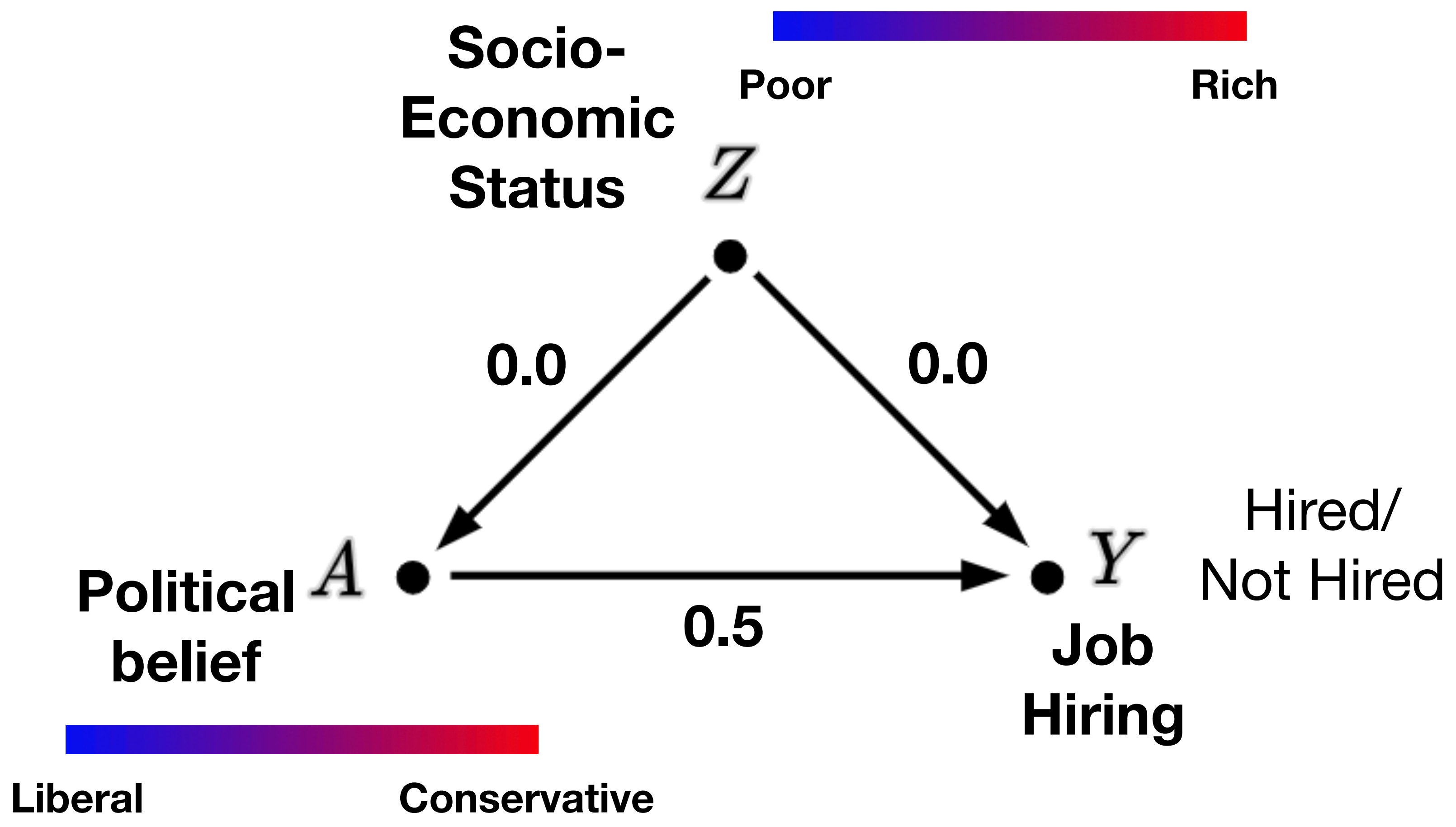


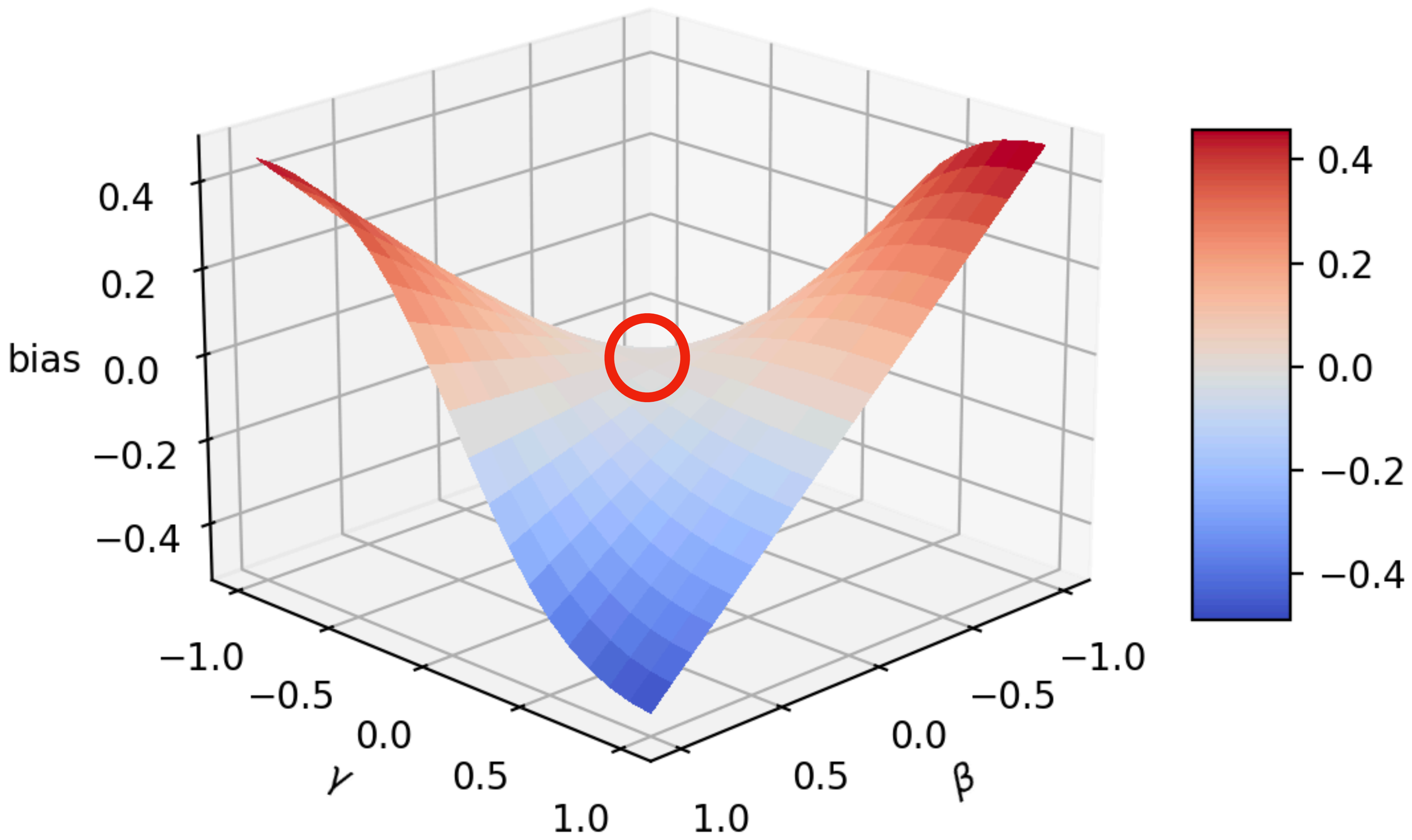


Confounding Bias = -0.4

Confounding Bias (Linear)

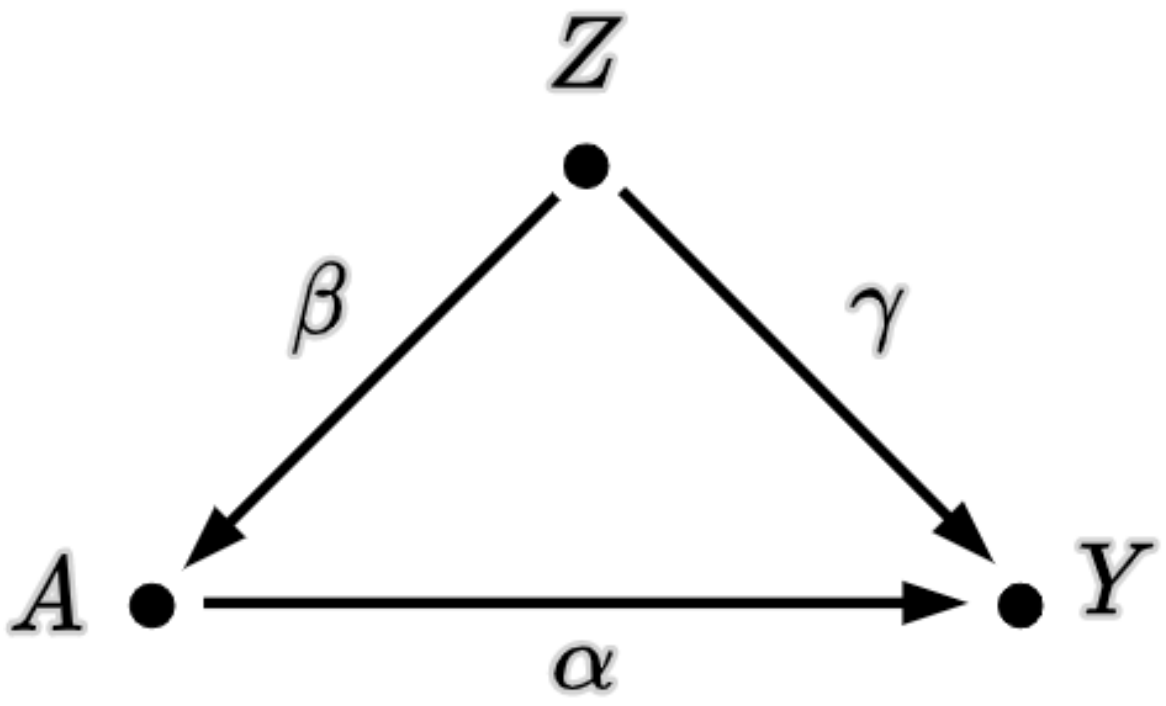


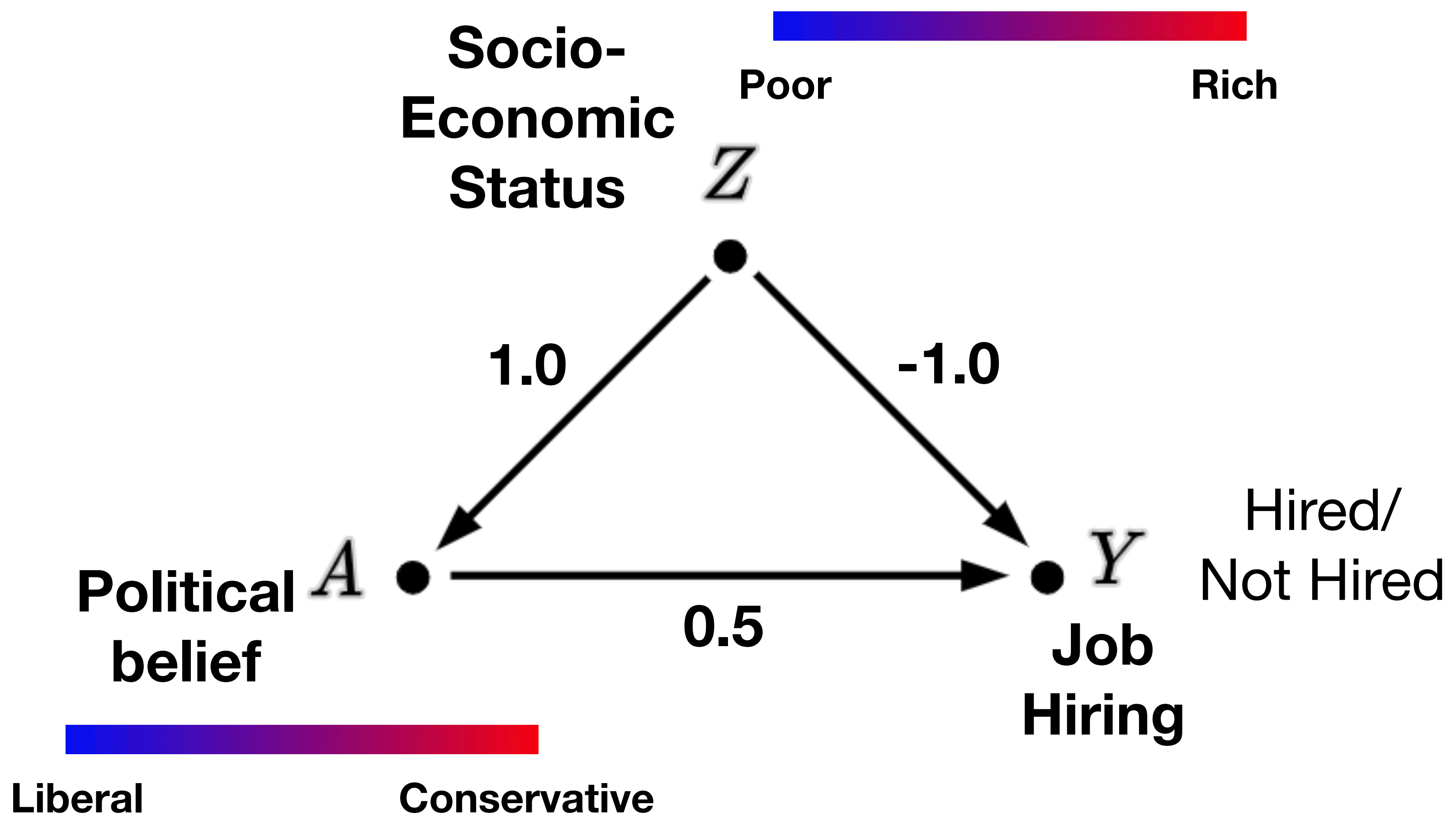


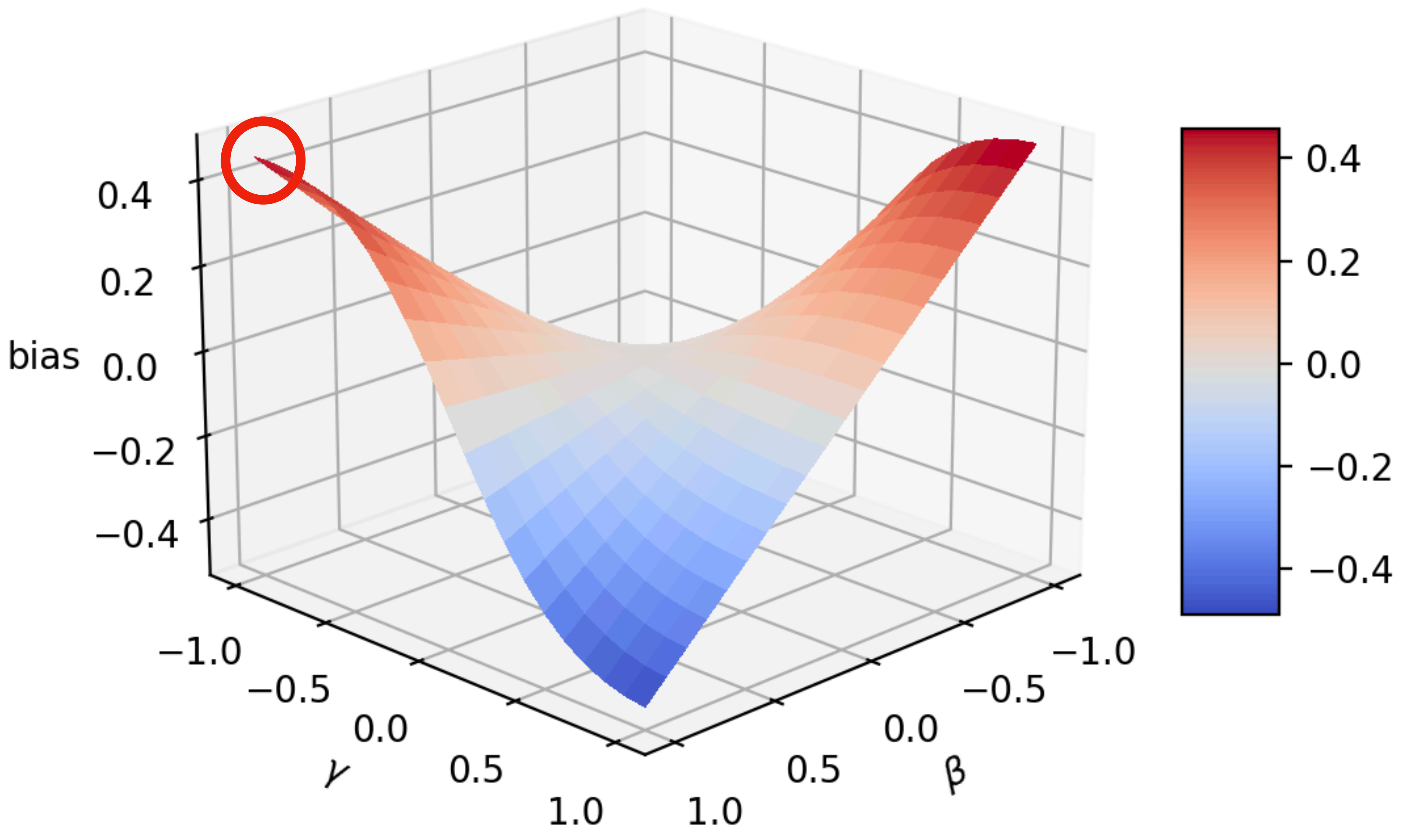


Confounding Bias = 0.0

Confounding Bias (Linear)

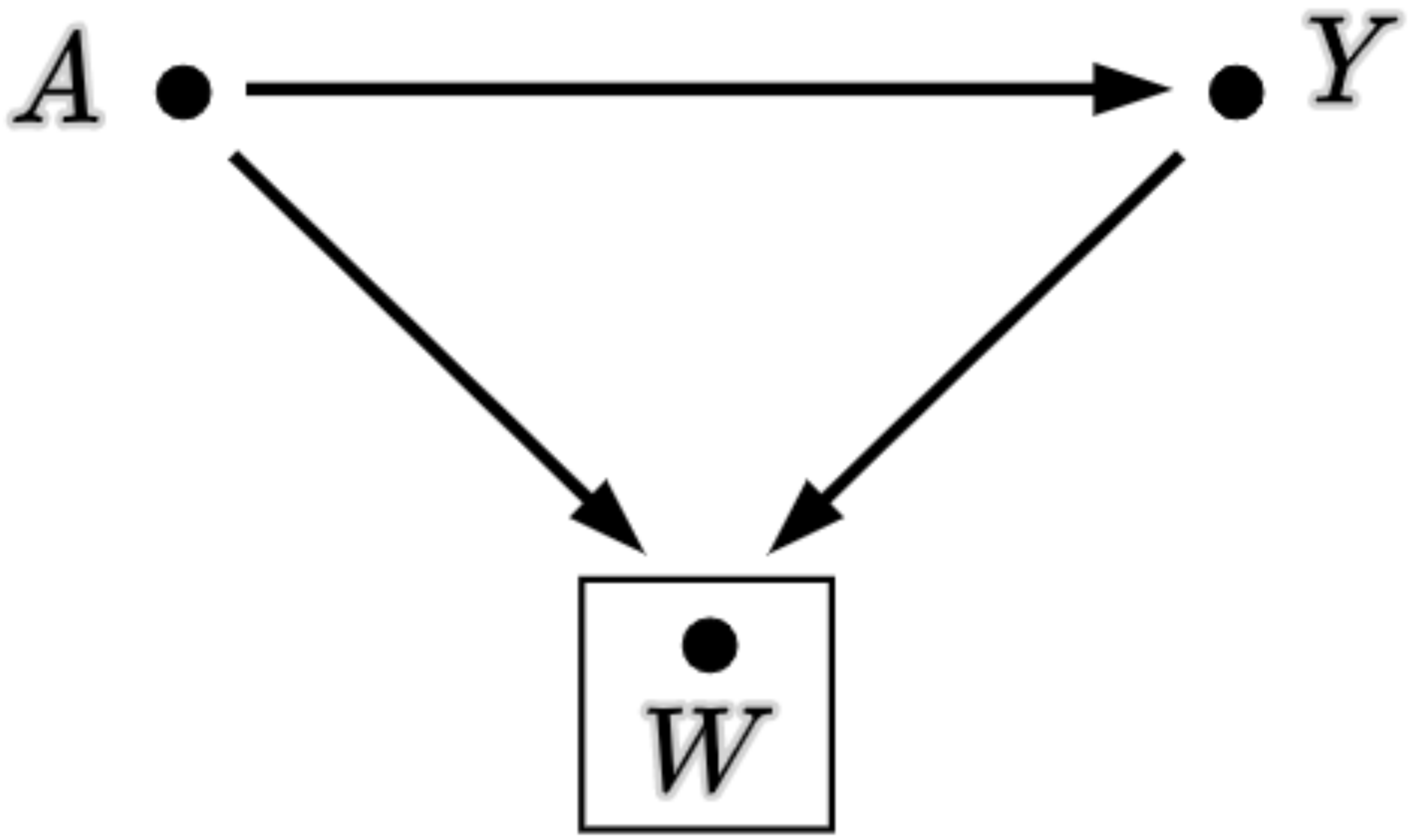


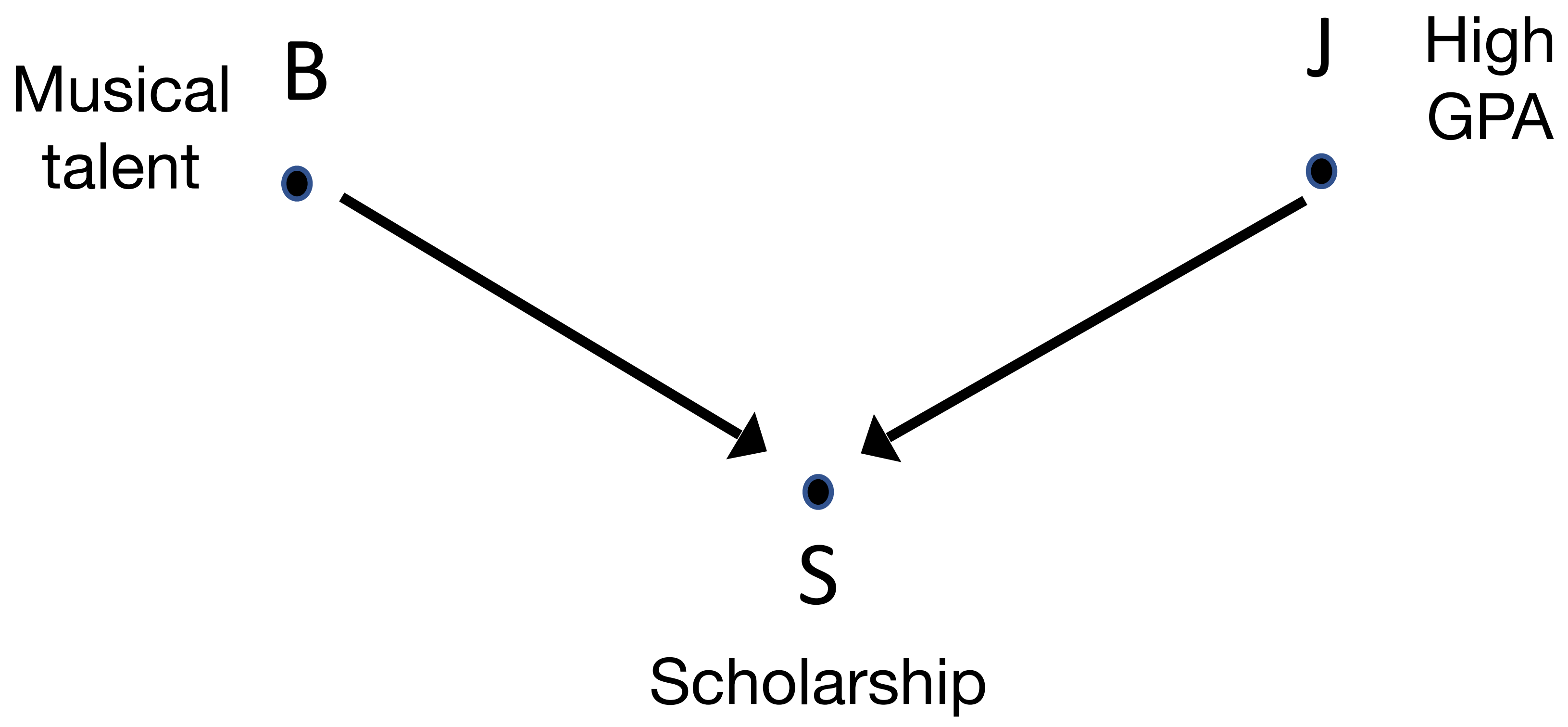




Confounding Bias = 0.4

Collider (Selection) Bias

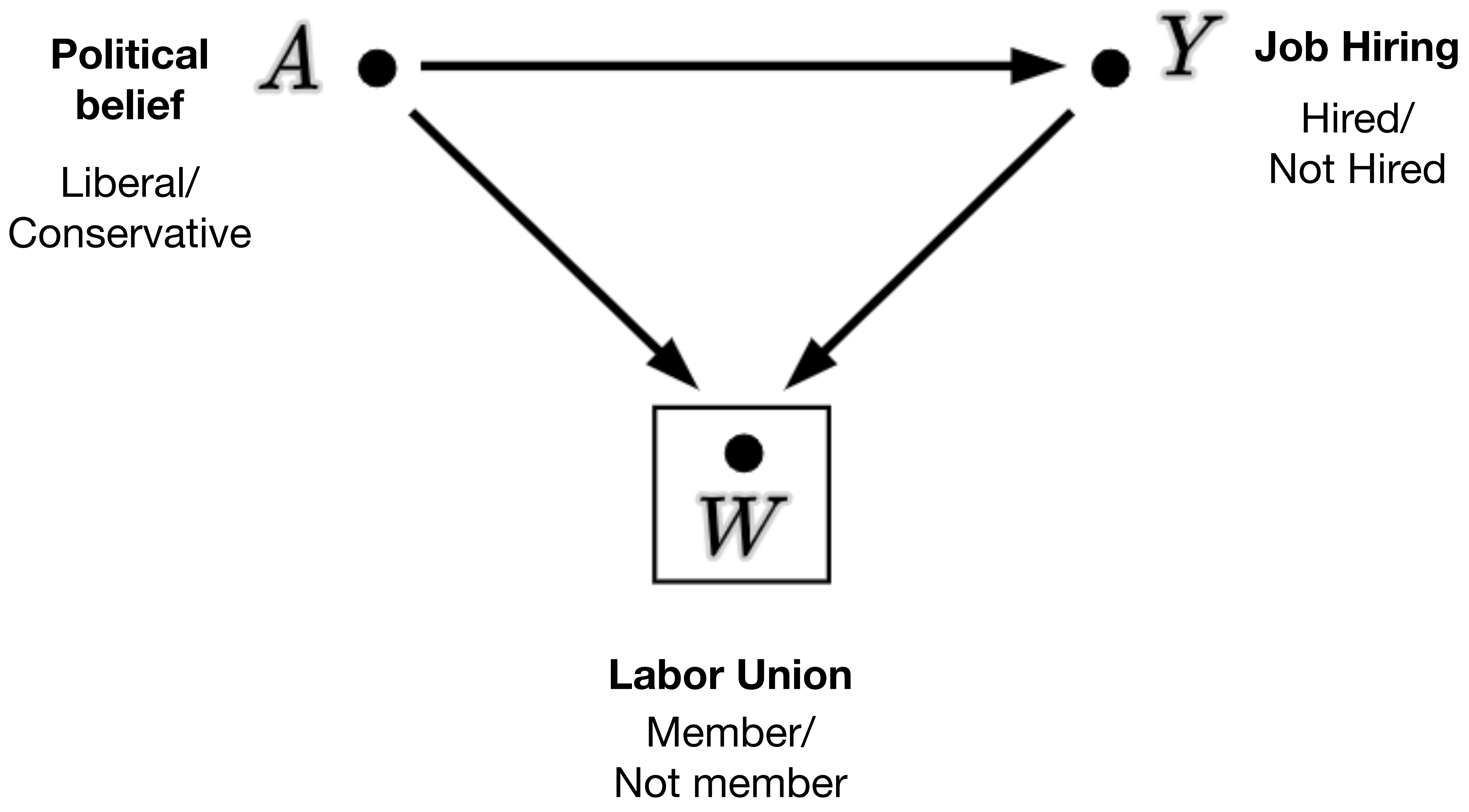




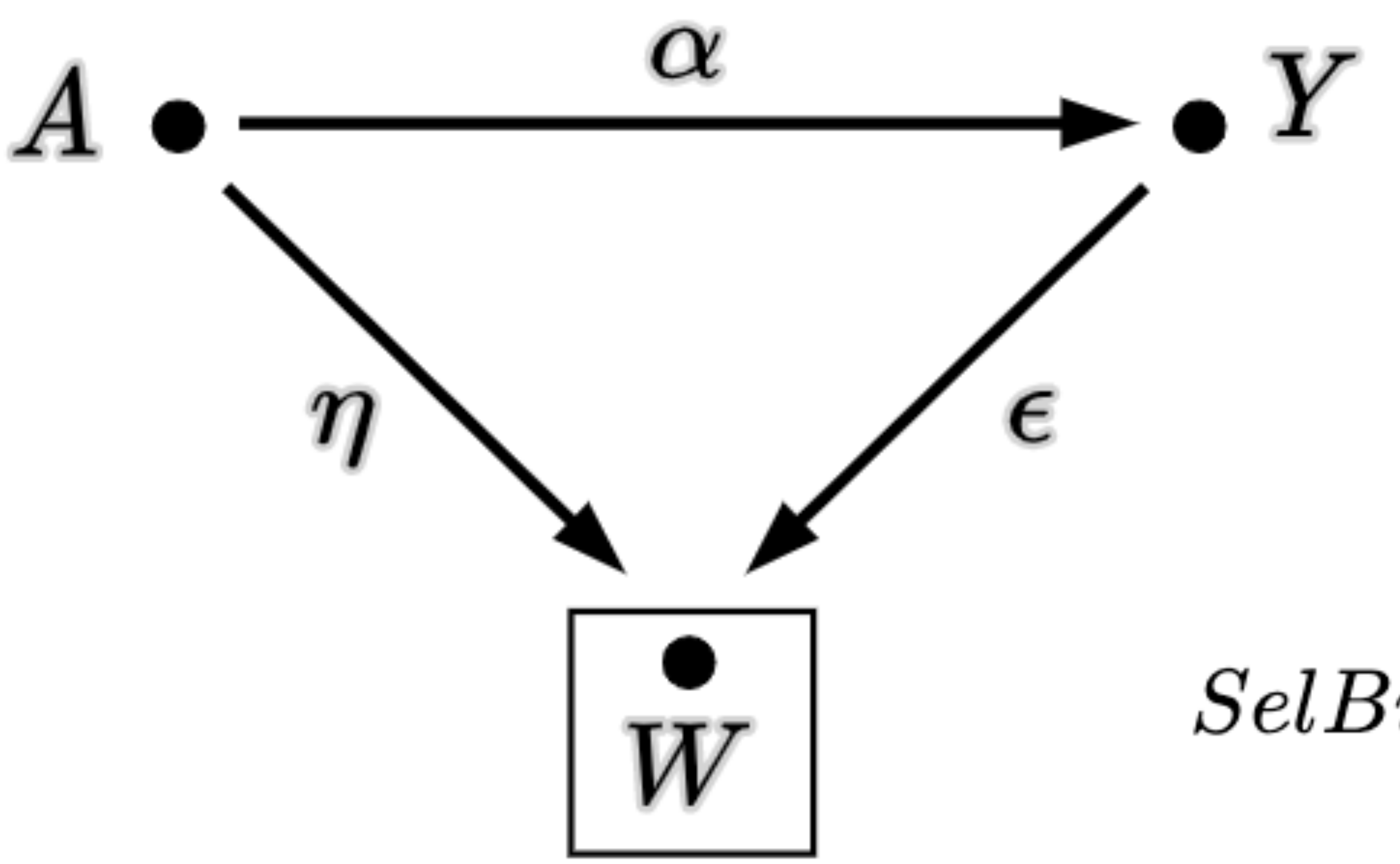
$B \perp J$

$B \not\perp J \mid S$

Collider (Selection) Bias



Collider (Selection) Bias (Linear Model)

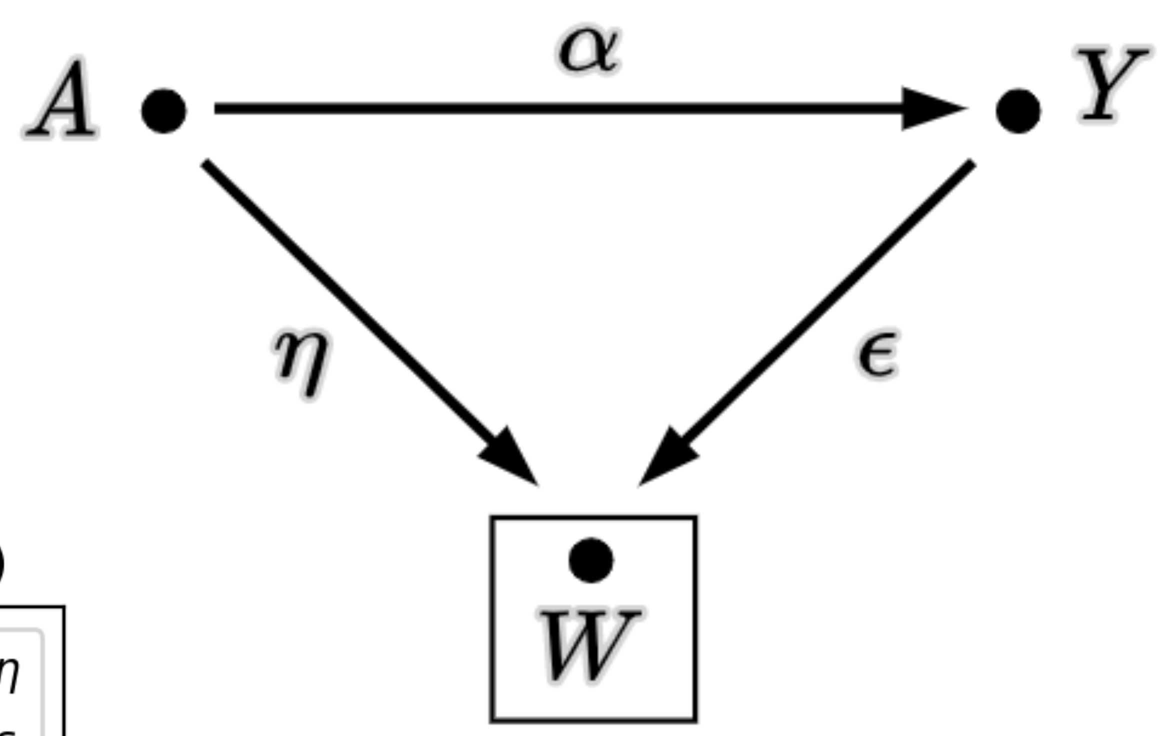


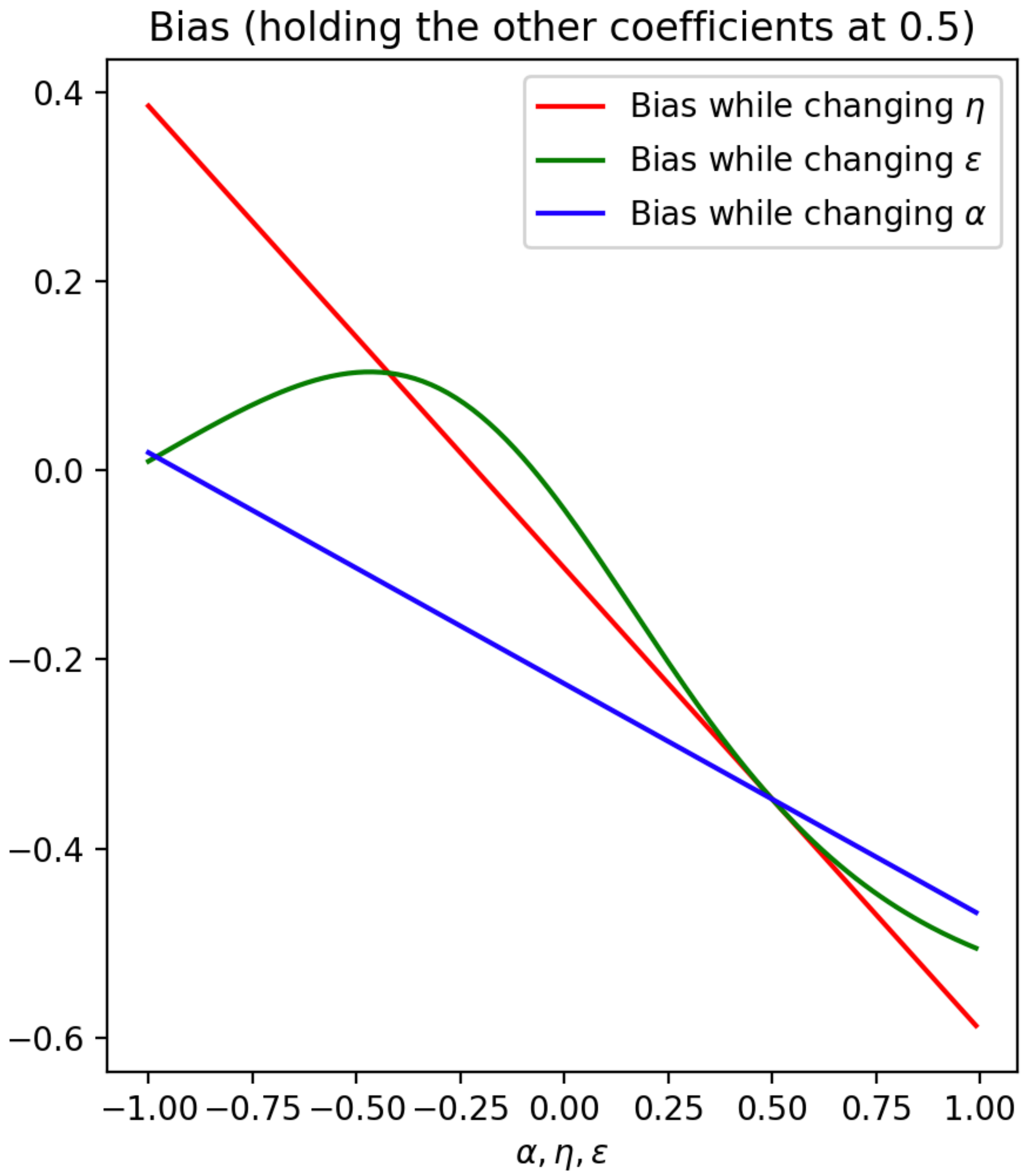
$$SelBias(Y,A) = \beta_{ya.w} - \beta_{ya}$$

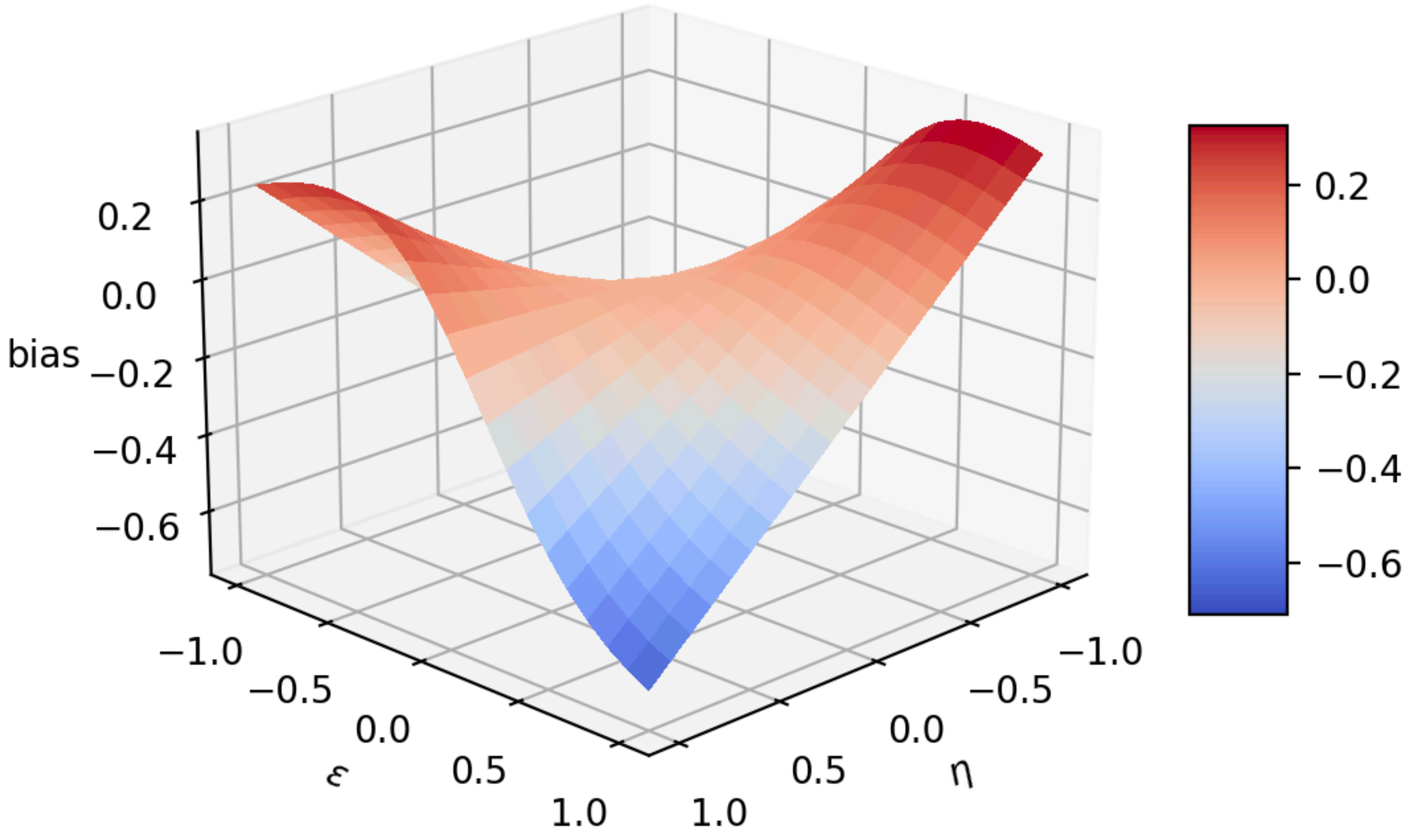
$$= \frac{\sigma_{aw}\left(\frac{\sigma_{ya}}{{\sigma_a}^2}\sigma_{aw} - \sigma_{yw}\right)}{{\sigma_a}^2{\sigma_w}^2 - {\sigma_{aw}}^2}$$

$$= \epsilon\frac{\alpha^2\eta + \alpha^3\epsilon{\sigma_a}^2 - \eta{\sigma_y}^2 - \alpha\epsilon{\sigma_y}^2}{{\sigma_w}^2 - {\sigma_a}^2(\eta+\alpha\epsilon)^2}$$

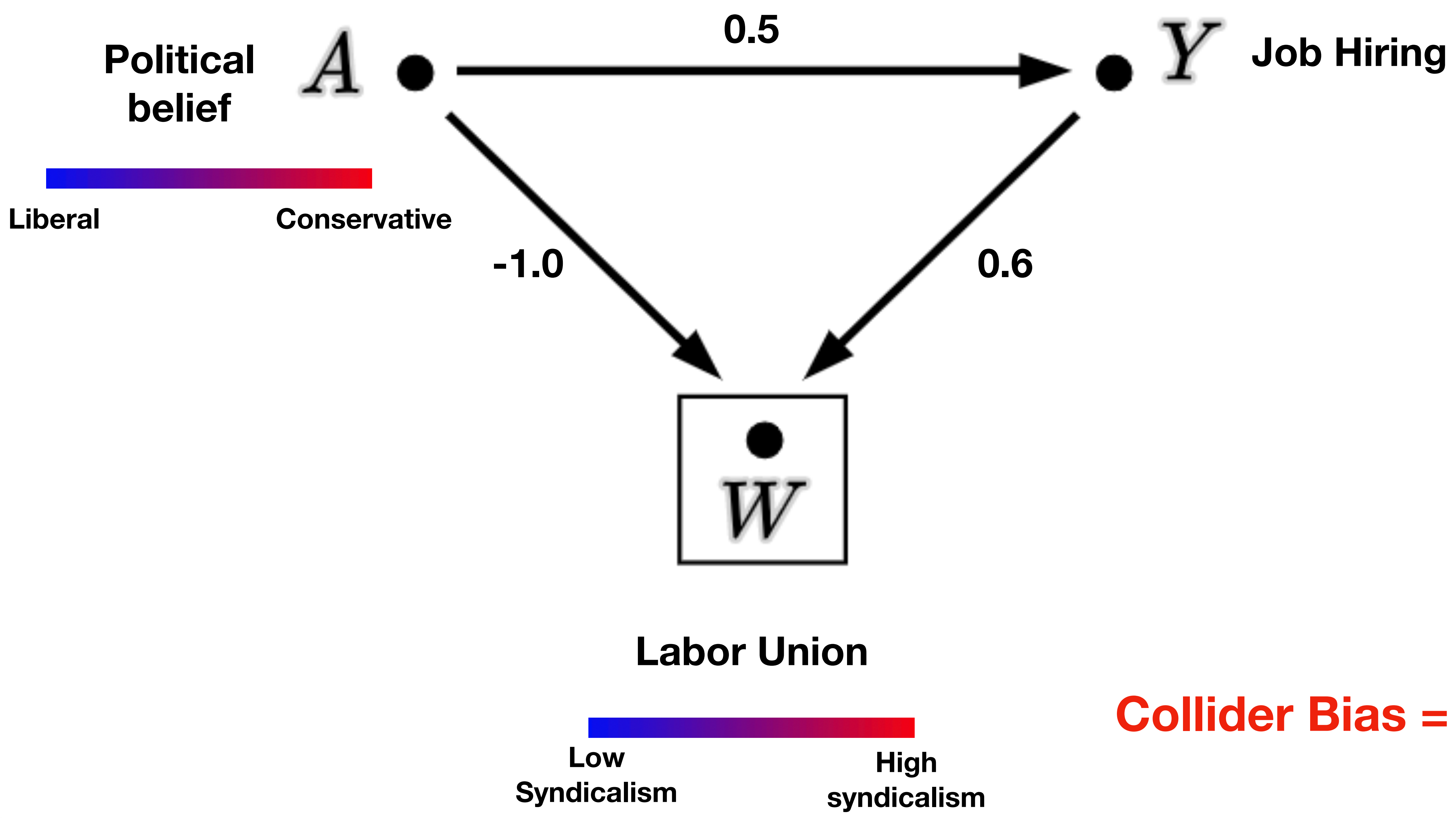
Collider (Selection) Bias (Linear Model)

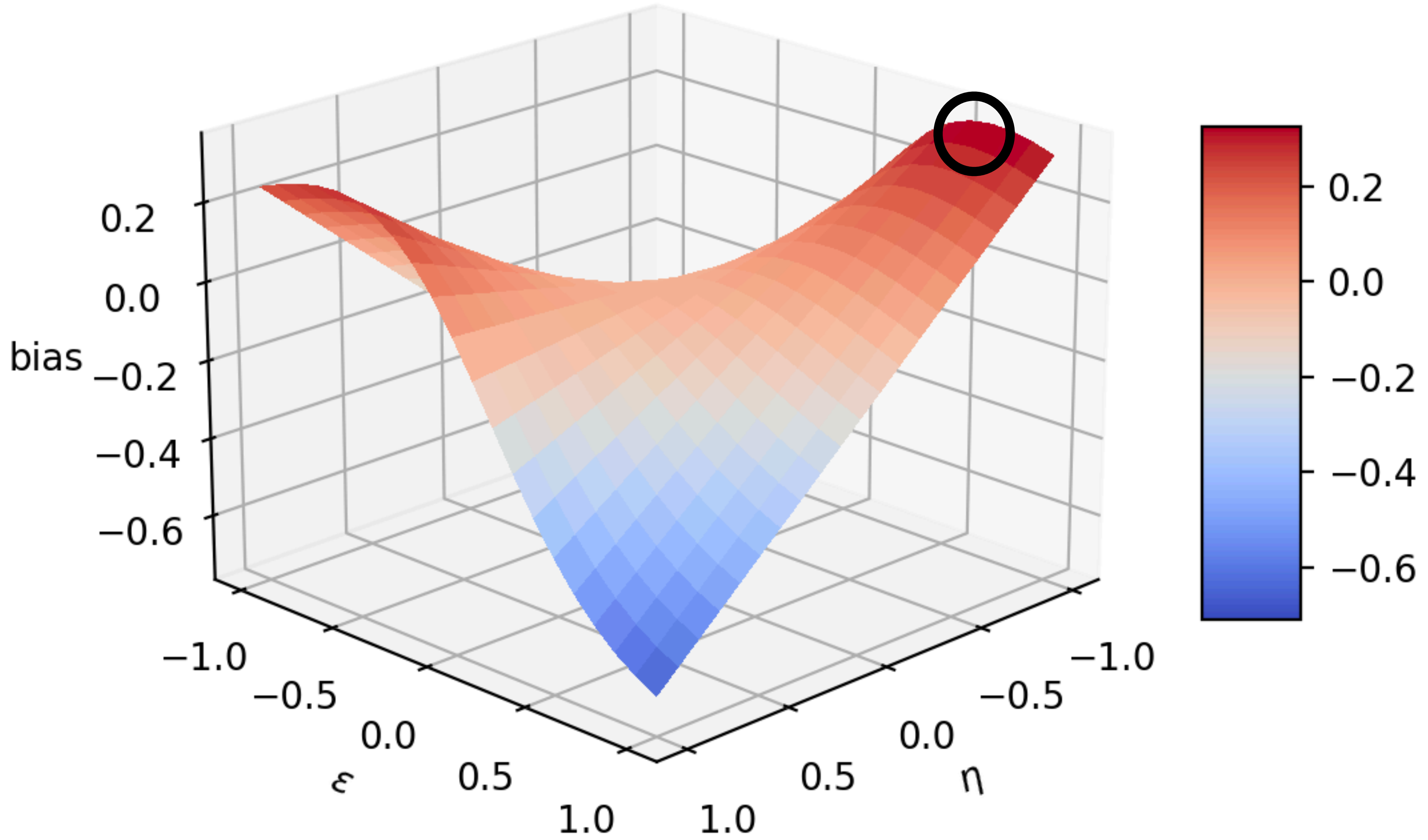




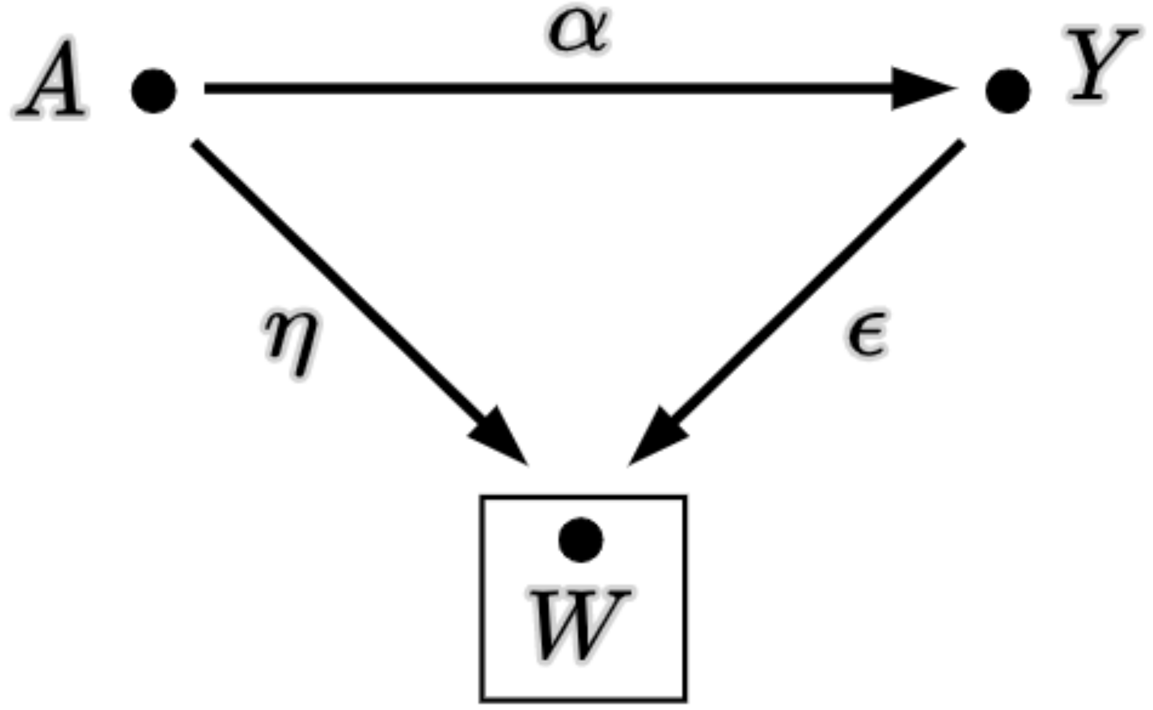


Collider (Selection) Bias (Linear Model)

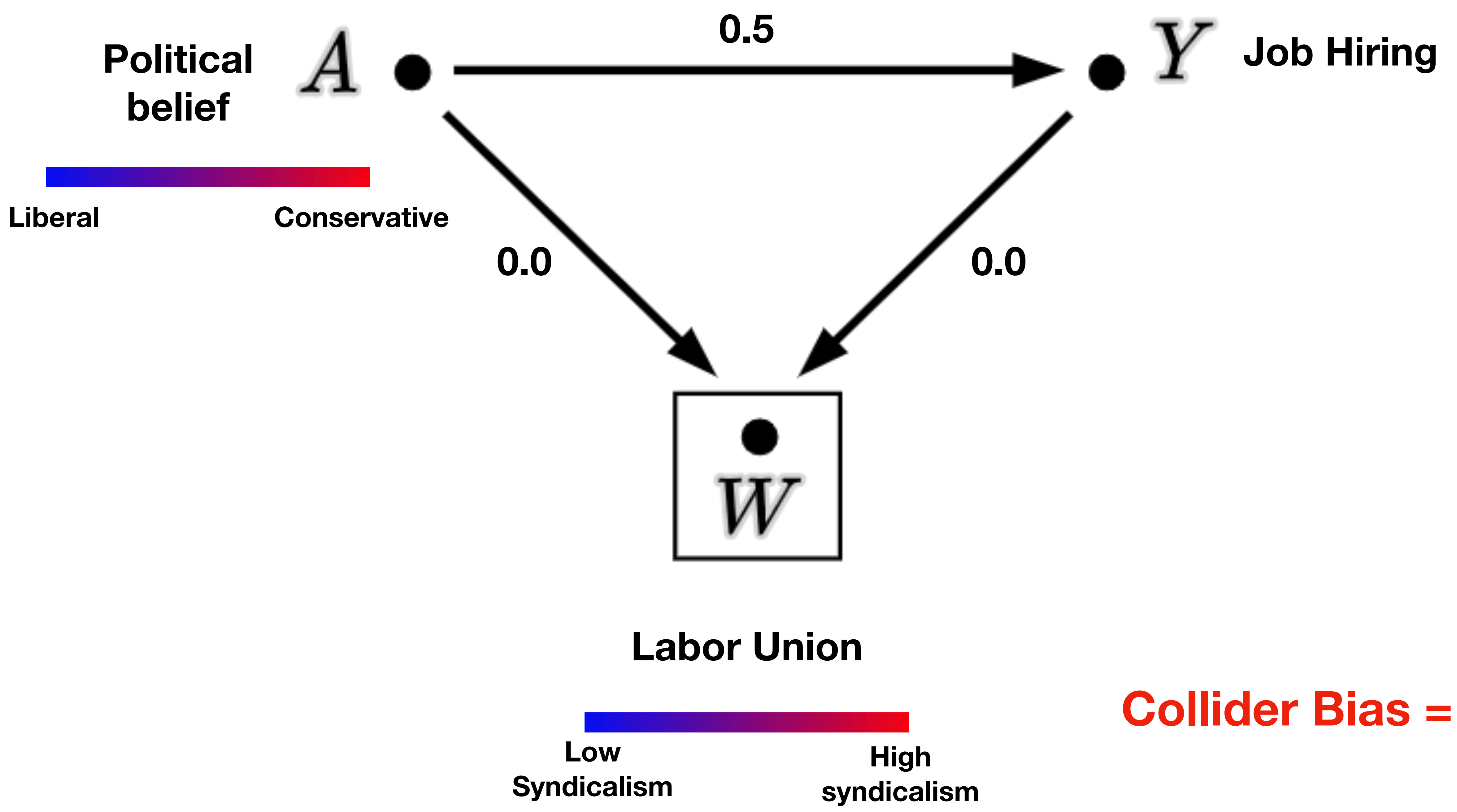


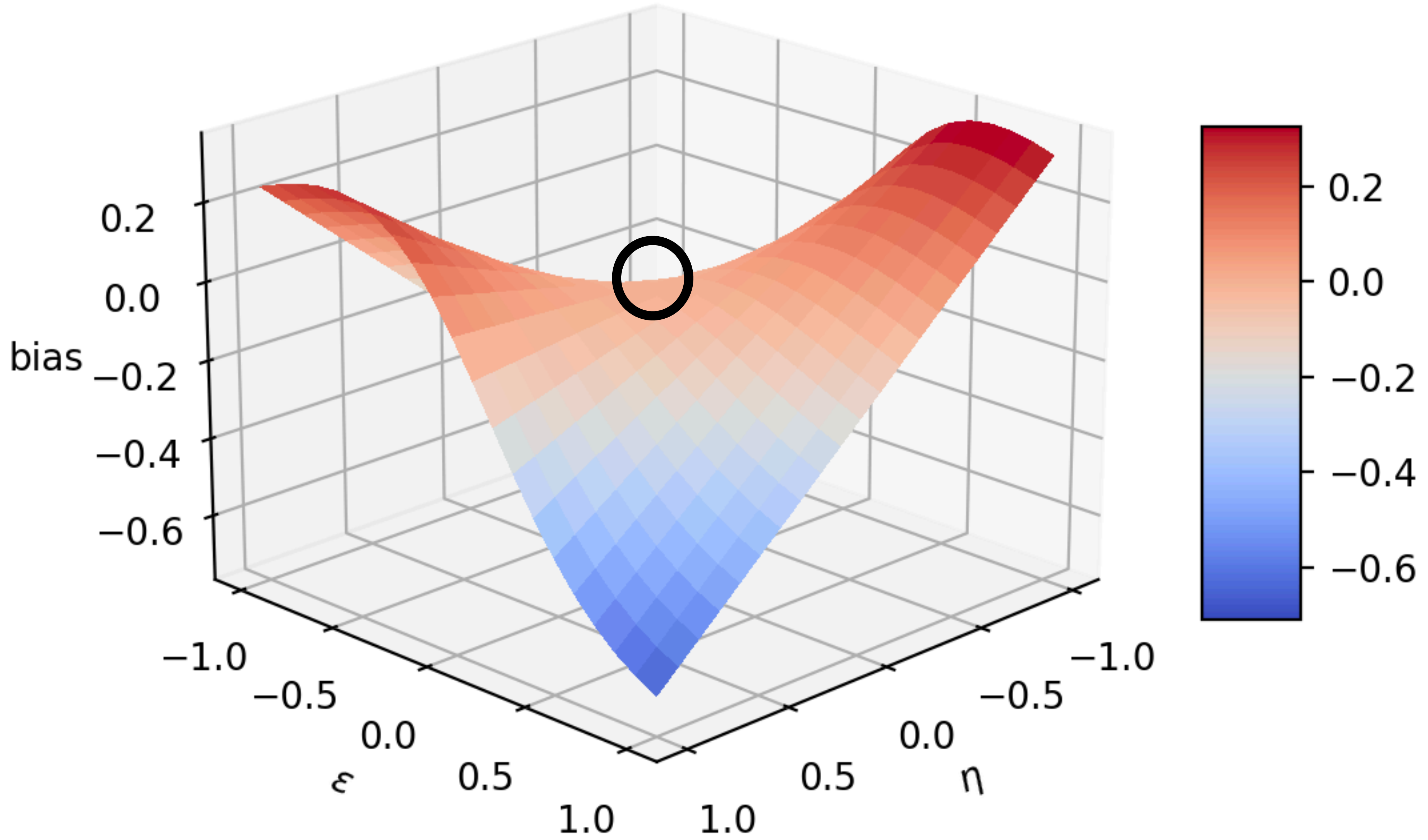


Collider Bias = 0.36

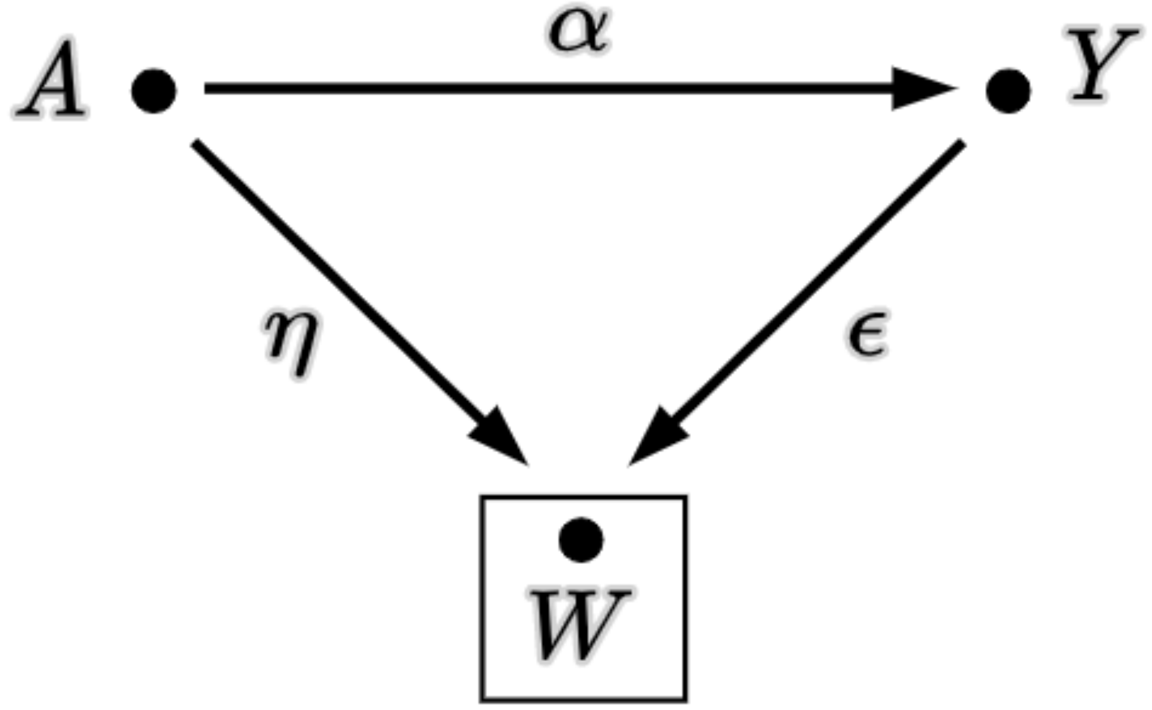


Collider (Selection) Bias (Linear Model)

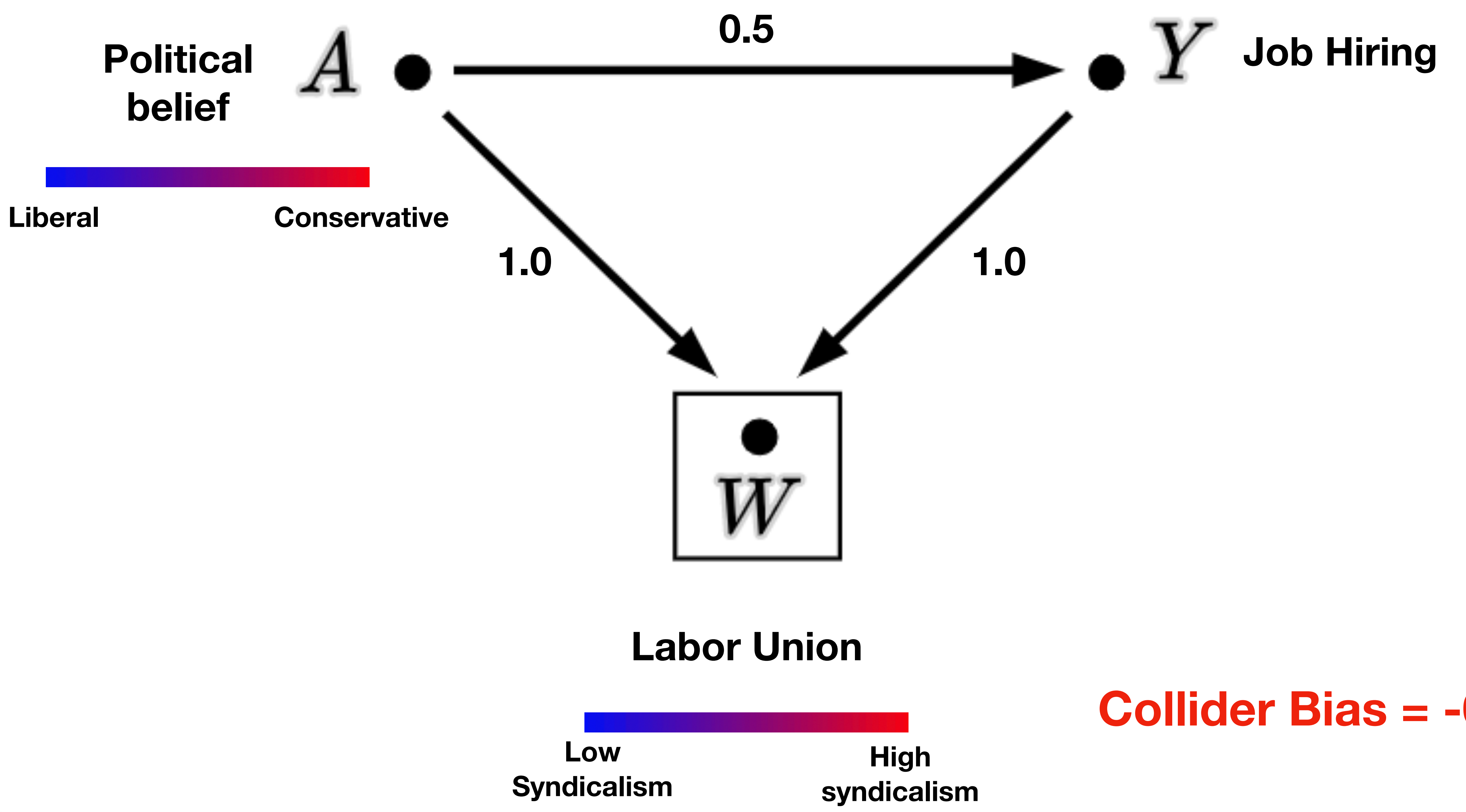


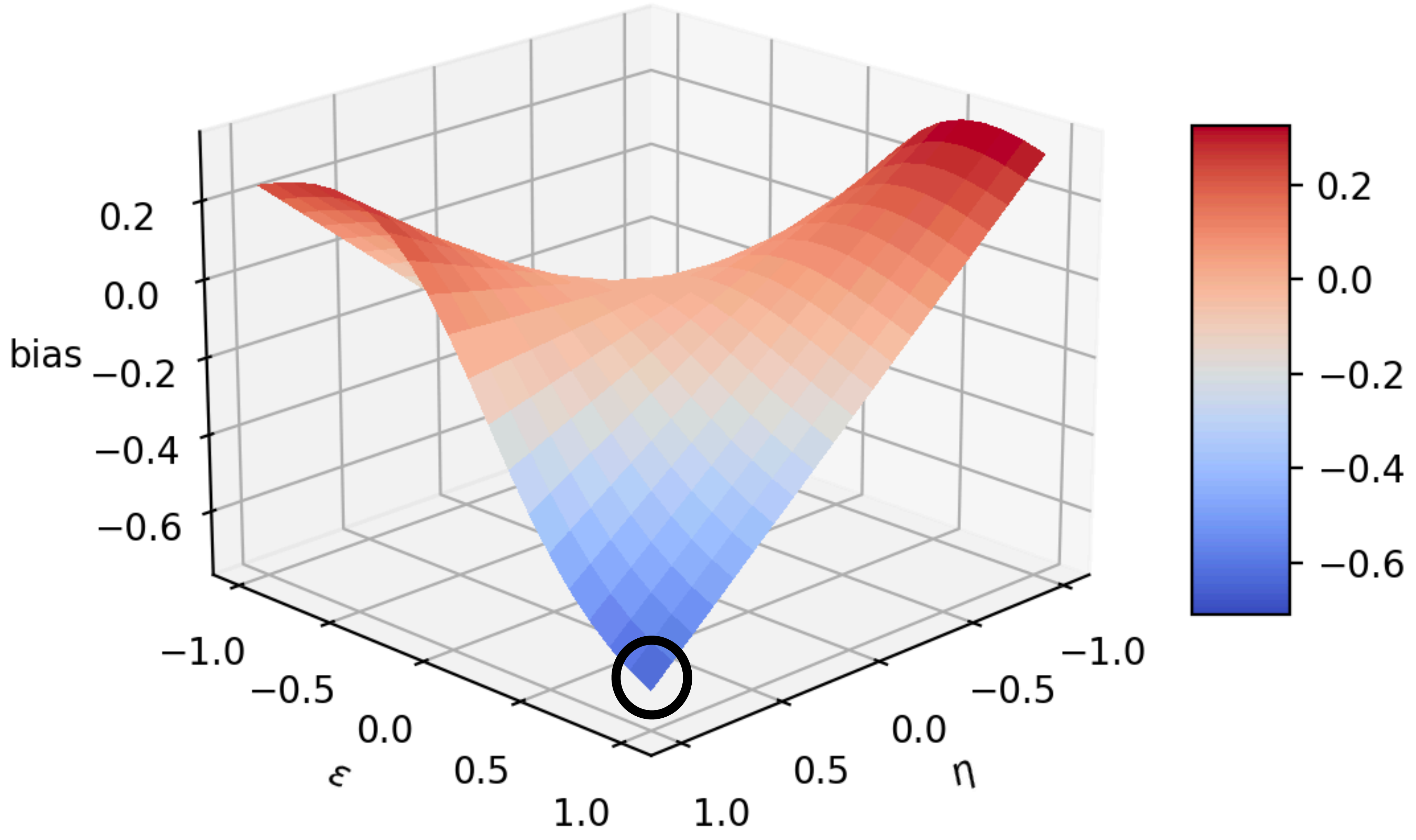


Collider Bias = 0.0

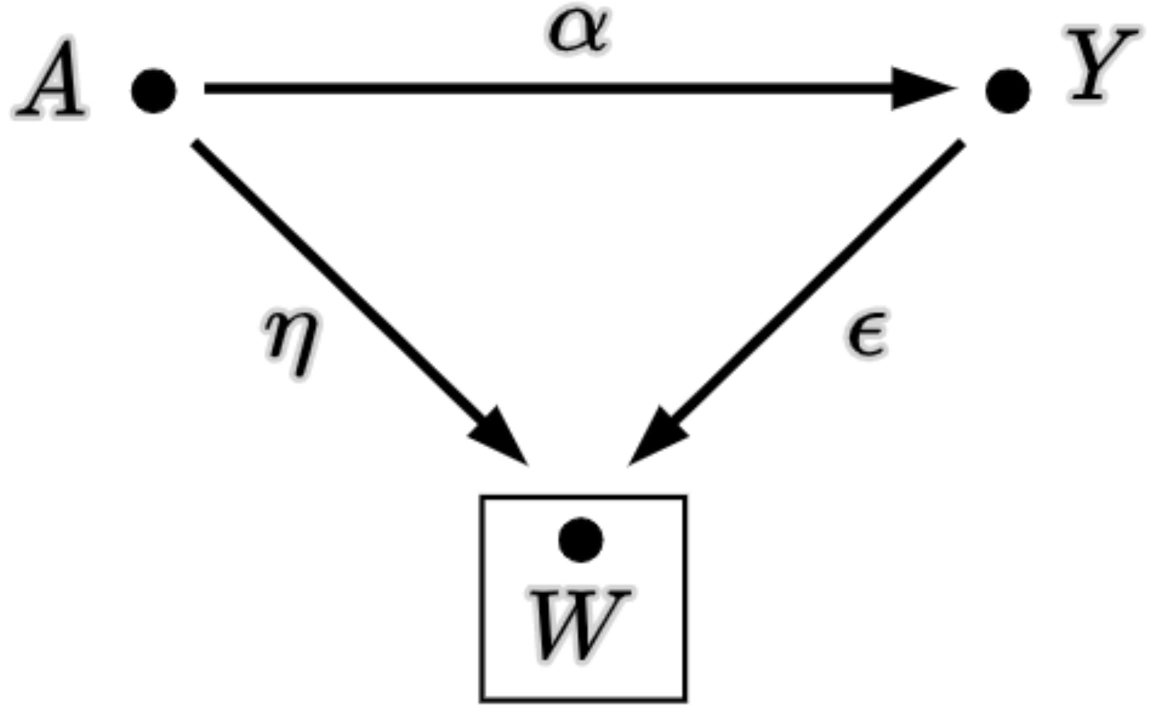


Collider (Selection) Bias (Linear Model)

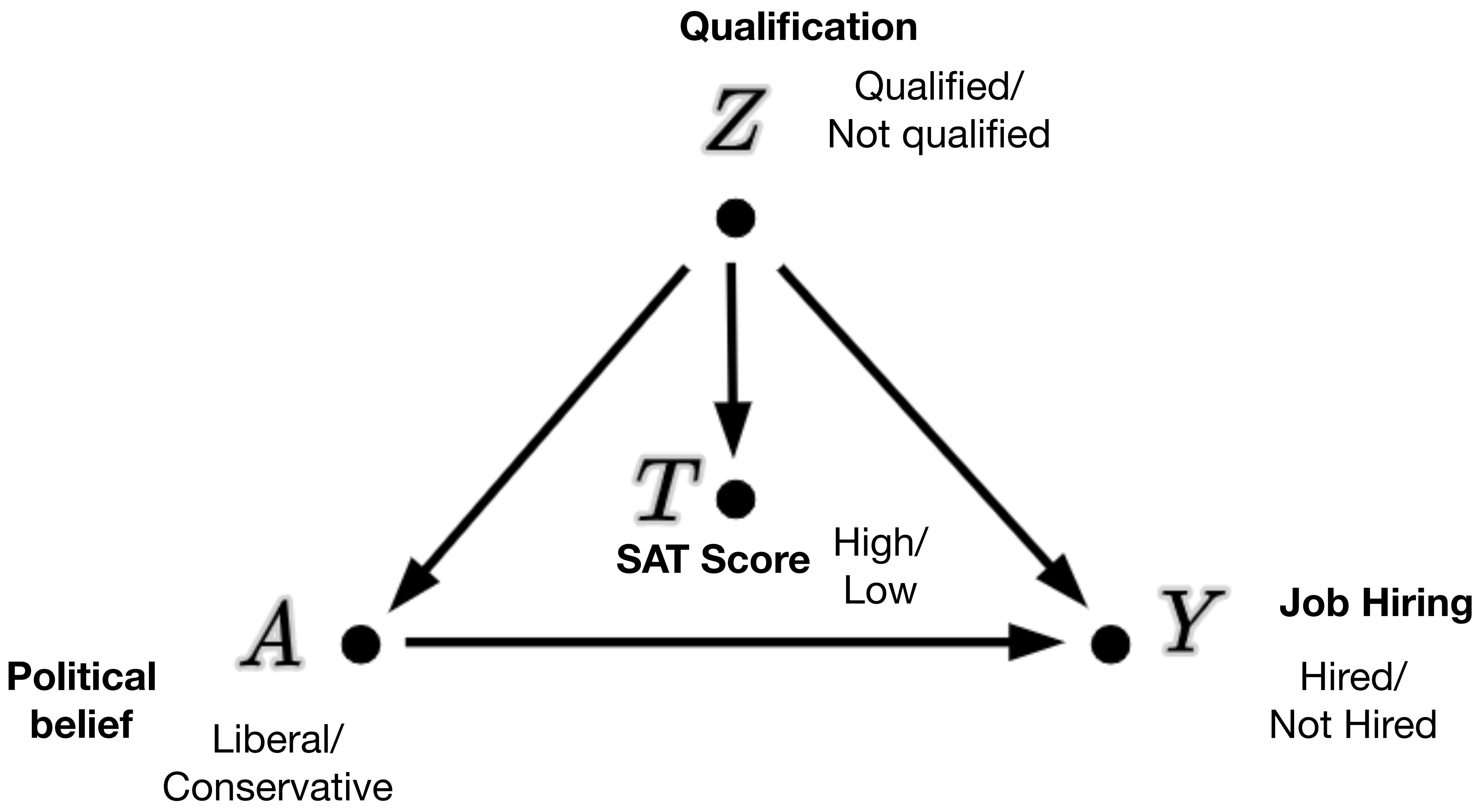




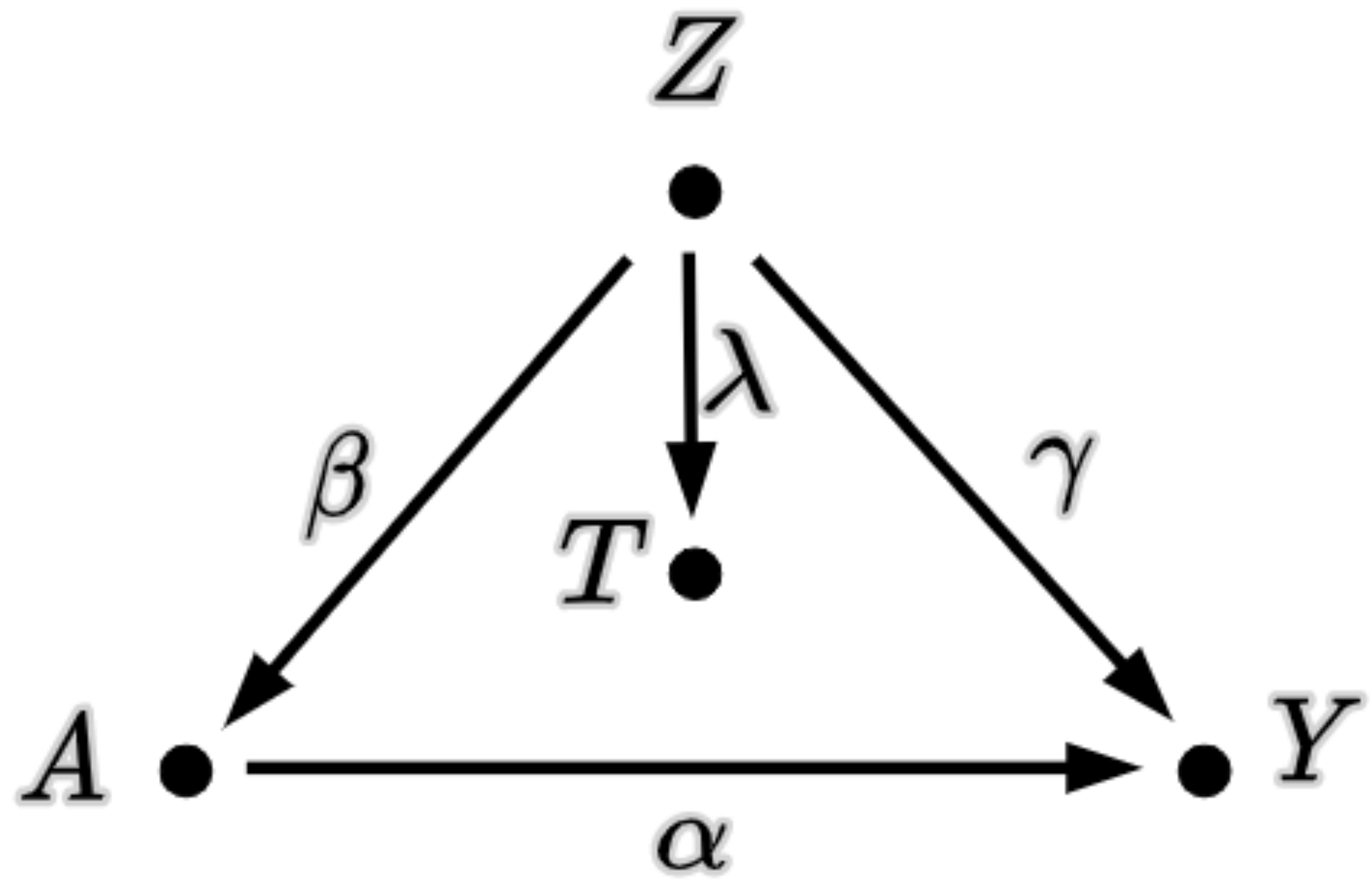
Collider Bias = -0.84



Measurement Bias

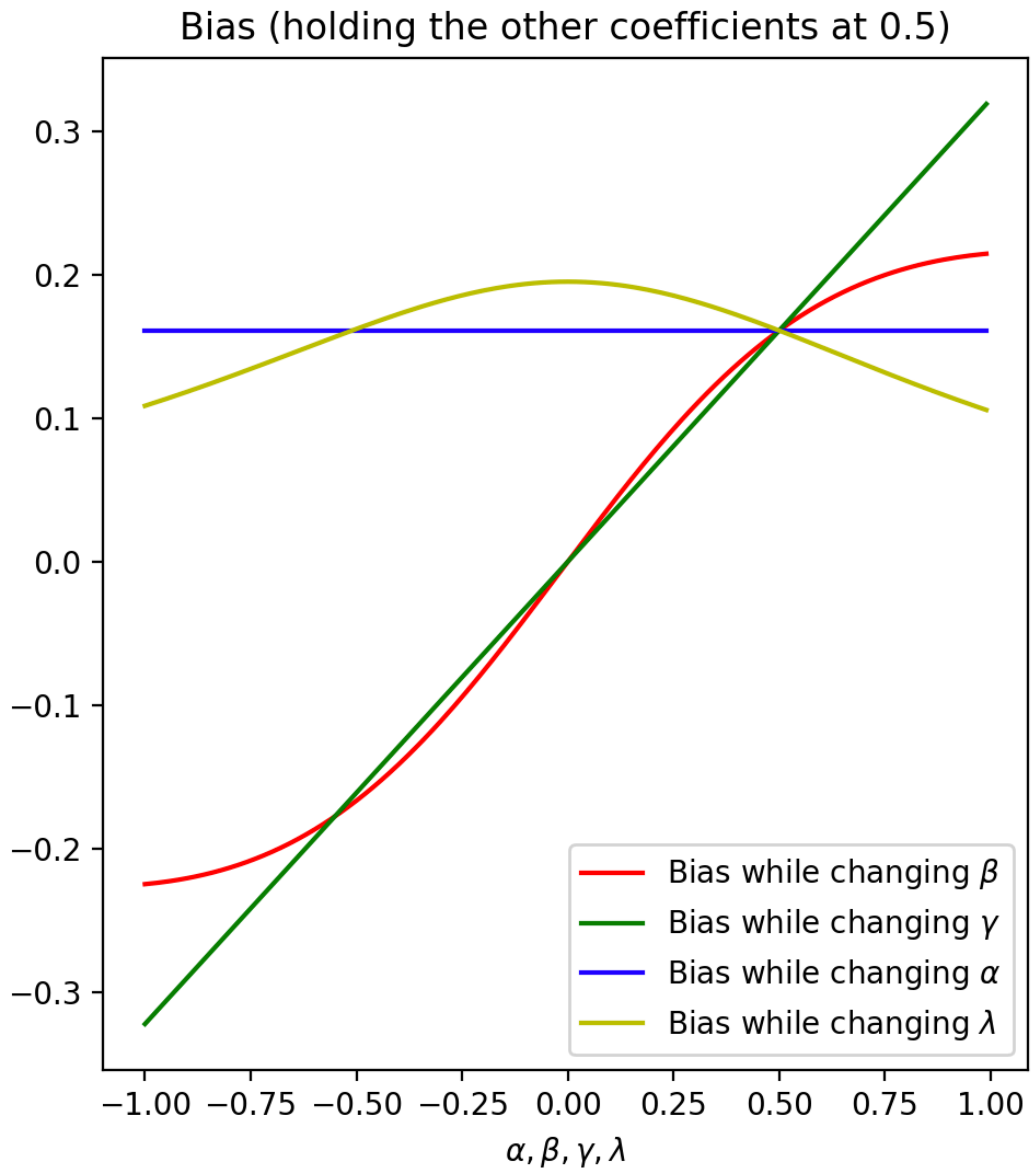


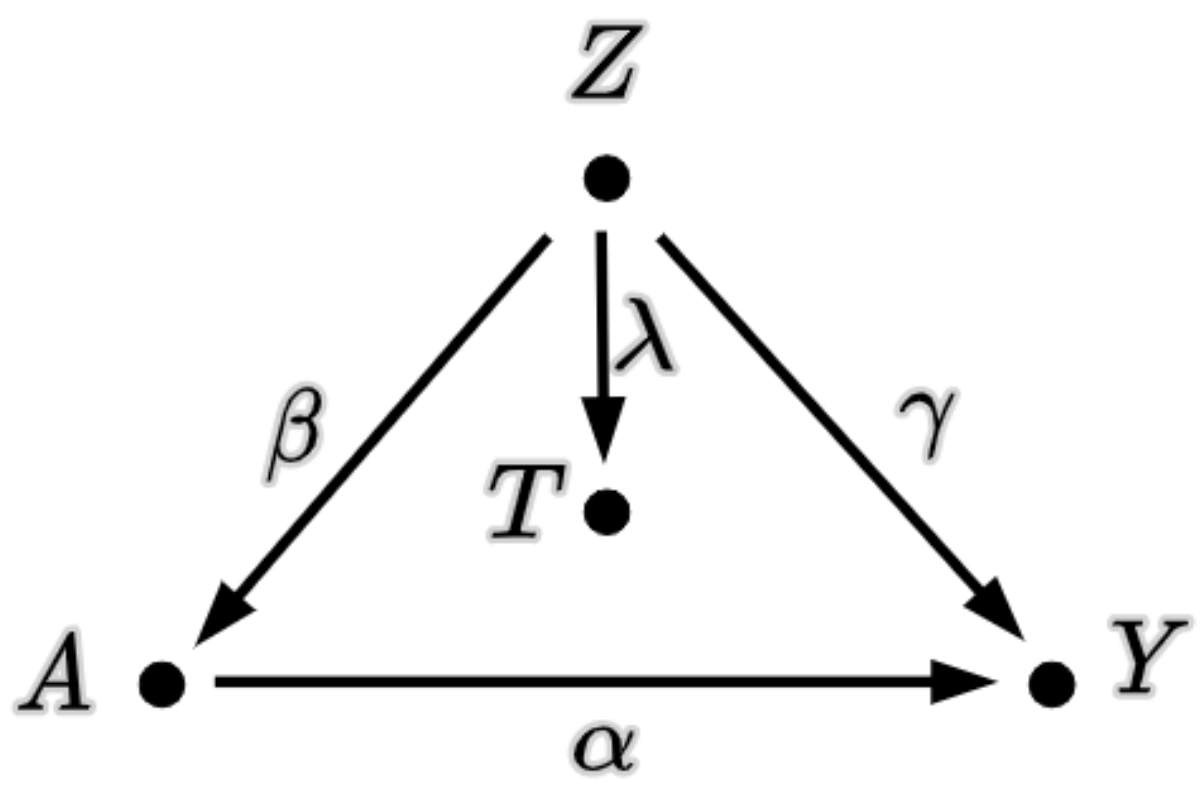
Measurement Bias (Linear Model)



$$
\begin{aligned}
MeasBias(Y, A) &= ACE(Y, A)_T - ACE(Y, A) \\
&= \beta_{ya.t} - \beta_{ya.z} \\
&= \frac{{\sigma_z}^2 \beta \gamma ({\sigma_t}^2 - {\sigma_z}^2 \lambda^2)}{{\sigma_a}^2 {\sigma_t}^2 - {\sigma_z}^4 \lambda^2 \beta^2}
\end{aligned}
$$

Measurement Bias (Linear Model)

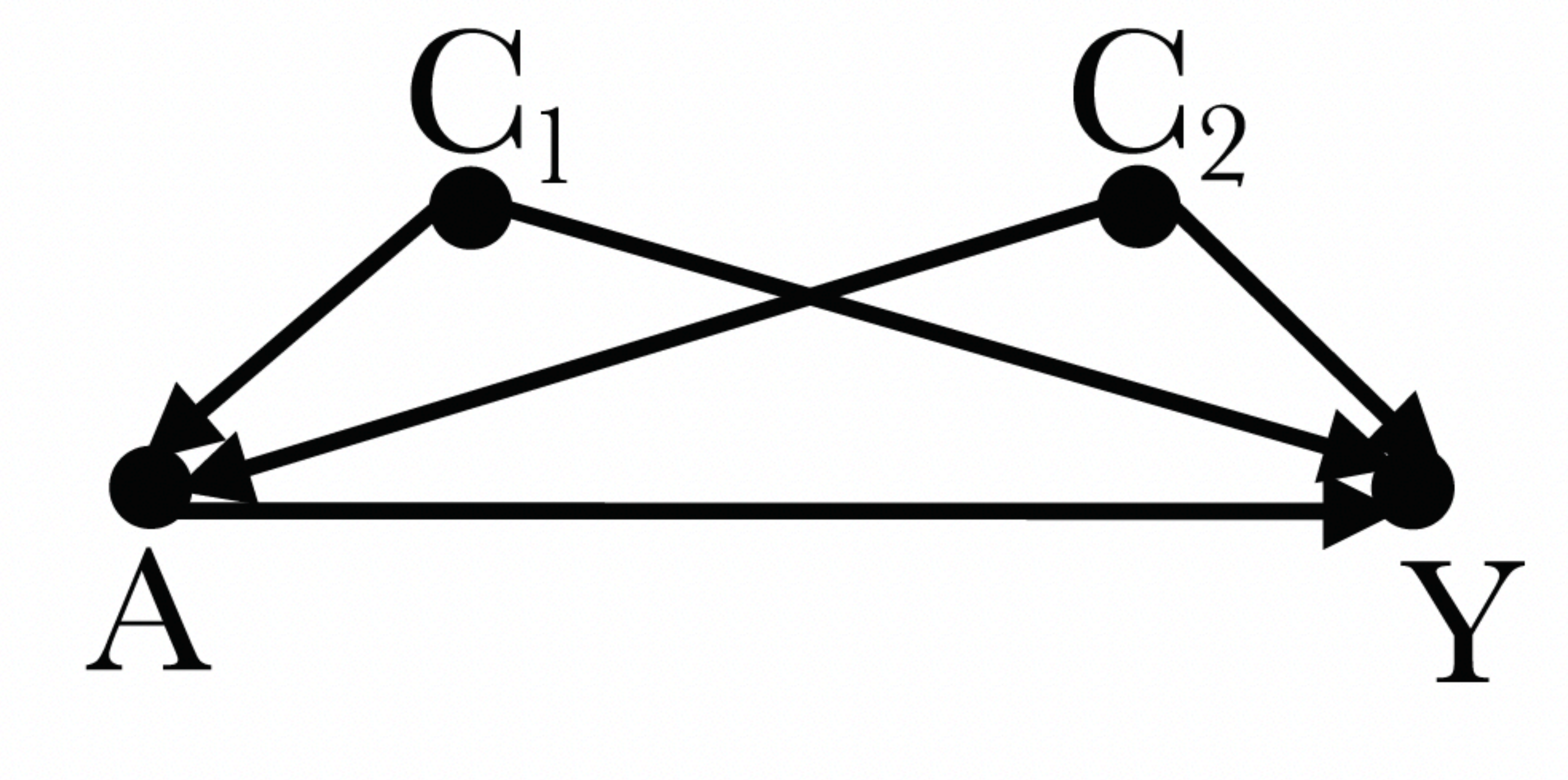


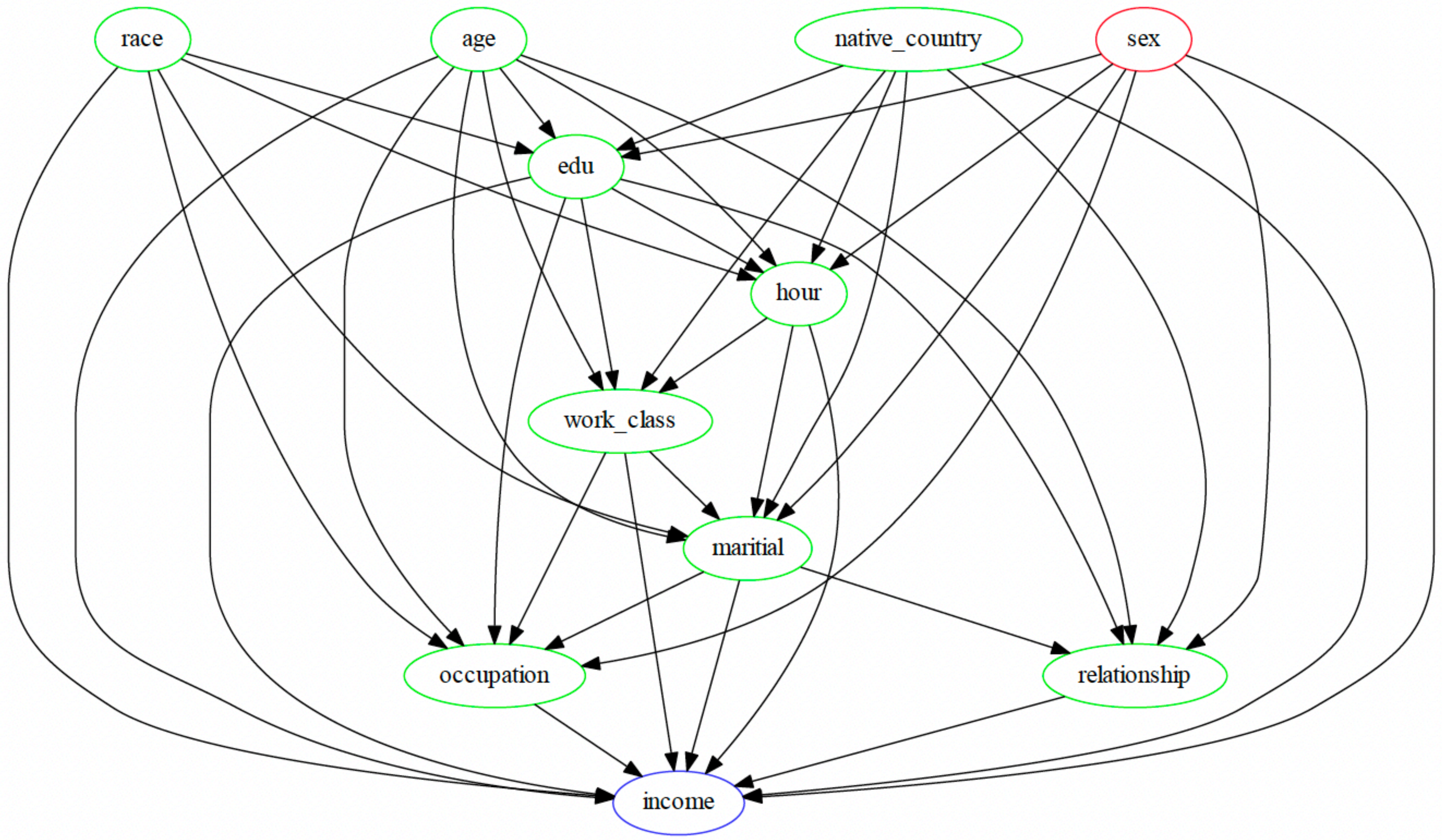


$$
\begin{aligned}
MeasBias(Y,A) &= ACE(Y,A)_T - ACE(Y,A) \\
&= \beta_{ya.t} - \beta_{ya.z} \\
&= \frac{{\sigma_z}^2 \beta\gamma({\sigma_t}^2 - {\sigma_z}^2\lambda^2)}{{\sigma_a}^2{\sigma_t}^2 - {\sigma_z}^4\lambda^2\beta^2}
\end{aligned}
$$

What’s next

- Understand more the magnitude of the bias in terms of the different model parameters.
- Quantify total bias in presence of several types of bias in the same setup
- Quantify bias in more complex causal models





Causal model of Adult benchmark dataset

Ethical AI sub-team at Comète

Catuscia Palamidessi
Director of research at Inria and leader of Comète team
catuscia@lix.polytechnique.fr

Frank Valencia
CNRS Researcher
frank.valencia@inria.fr

Sami Zhioua
Advanced researcher at Inria, LIX, École Polytechnique
sami.zhioua@lix.polytechnique.fr

Ruta Binkyte
PhD student at Inria, LIX, École Polytechnique
ruta.binkyte-sadauskiene@inria.fr

Mario Alvim
Researcher
mario.ferreira-alvim-junior@inria.fr>

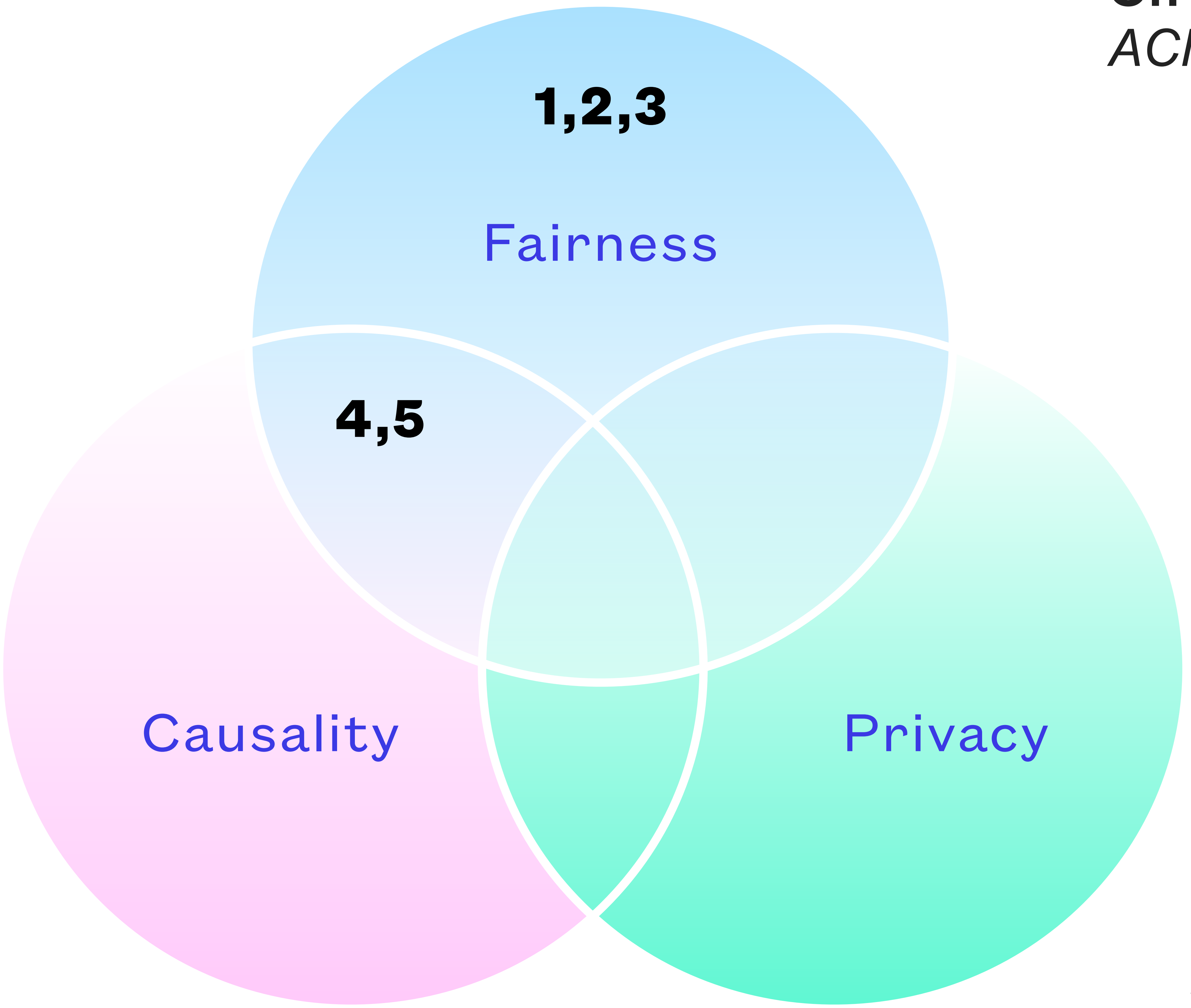
Karima Makhlouf
PhD student at Inria, LIX, École Polytechnique
karima.makhlouf@lix.polytechnique.fr

Carlos Pinzón
PhD student at Inria, LIX, École Polytechnique
carlos.pinzon@inria.fr

Héber H. Arcolezi
Posdoctoral Researcher at Inria, LIX, École Polytechnique
heber.hwang-arcolezi@inria.fr

Szilvia Lestyan
Postdoctoral researcher

Completed



1. Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021).
Machine learning fairness notions: Bridging the gap with real-world applications.
Information Processing & Management Journal.

2. Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021).
On the applicability of machine learning fairness notions.
ACM SIGKDD Explorations Newsletter.

3. Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020).
Survey on causal-based machine learning fairness notions.
Under review.

4. Pinzón, C., Palamidessi, C., Piantanida, P., & Valencia, F. (2022, June).
On the Impossibility of Non-trivial Accuracy in Presence of Fairness Constraints.
In *Proceedings of the AAAI Conference on Artificial Intelligence*.

5. Binkytė, R., Makhlouf, K., Pinzón, C., Zhioua, S., & Palamidessi, C.
Causal Discovery for Fairness.
NeurIPS 2022.
Workshop on Algorithmic Fairness through the Lens of Causality and Privacy.

Causal Discovery for Fairness

Rūta Binkytė-Sadauskienė
ruta.binkyte-sadauskiene@inria.fr
INRIA, École Polytechnique, IPP
Paris, France

Karima Makhlouf
karima.makhlouf@lix.polytechnique.fr
INRIA, École Polytechnique, IPP
Paris, France

Carlos Pinzón
carlos.pinzon@inria.fr
Inria, École Polytechnique, IPP
Paris, France

Sami Zhioua
sami.zhioua@lix.polytechnique.fr
INRIA, École Polytechnique, IPP
Paris, France

Catuscia Palamidessi
catuscia@lix.polytechnique.fr
Inria, École Polytechnique, IPP
Paris, France

ABSTRACT

It is crucial to consider the social and ethical consequences of AI and ML based decisions for the safe and acceptable use of these emerging technologies. Fairness, in particular, guarantees that the ML decisions do not result in discrimination against individuals or minorities. Identifying and measuring reliably fairness/discrimination is better achieved using causality which considers the causal relation, beyond mere association, between the sensitive attribute (e.g.

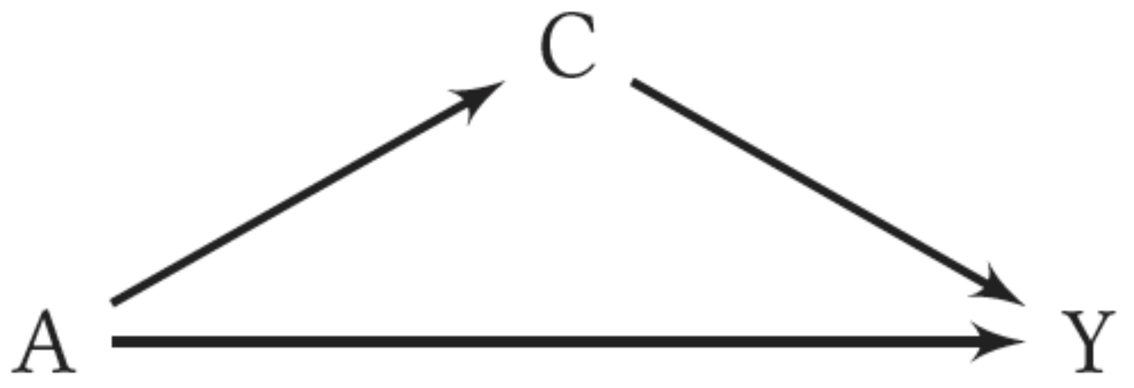
criteria have been introduced in the literature to assess discrimination (statistical parity [13], equal opportunity [21], calibration [12], etc.) [42]. The most recent fairness criteria, however, are causal-based [40] and reflect the now widely accepted idea that causality is necessary to appropriately address the problem of fairness. There are at least three benefits of using causality to assess fairness. First, in presence of a common cause (confounder) between the sensitive attribute A (e.g. gender) and the decision Y (e.g. job hiring),

* NeurIPS 2022 Workshop on Algorithmic Fairness through the Lens of Causality and Privacy
* Long version available at arxiv: https://arxiv.org/abs/2206.06685

Causal Discovery for Fairness

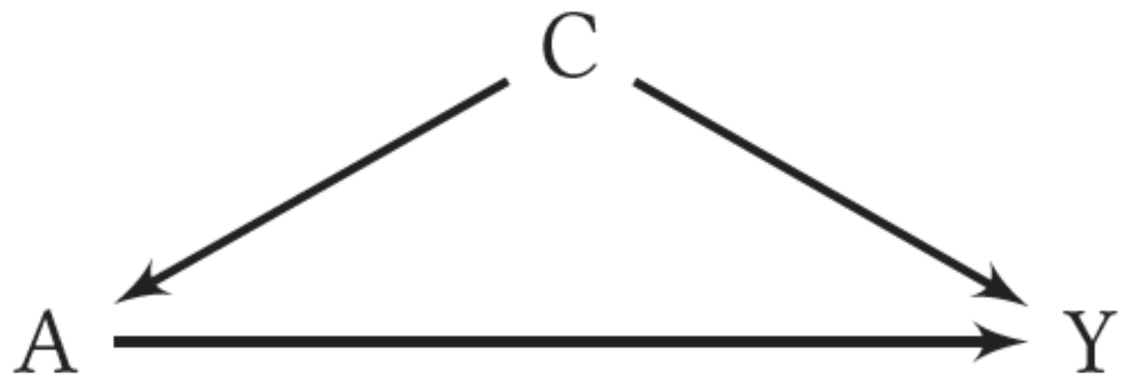
Different causal discovery algorithms (PC, FCI, GES, LiNGAM, etc.) may lead to different causal graphs.

We show that even slight differences in causal graphs can have significant impact on fairness conclusions.



$$TE_{a_1,a_0}(y^+) = \mathbb{P}(y^+_{a_1}) - \mathbb{P}(y^+_{a_0})$$
$$= \mathbb{P}(y^+|A = a_1) - \mathbb{P}(y^+|A = a_0).$$

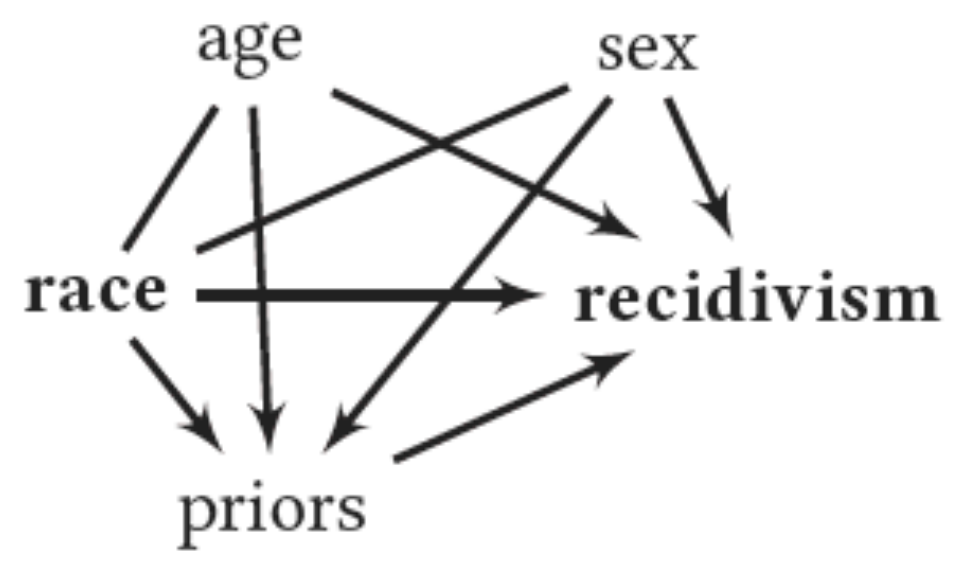
$$NIE_{a_1,a_0}(y^+) = \sum_{c \in dom(C)} \mathbb{P}(Y = y^+|A = a_0, C = c)$$
$$(\mathbb{P}(C = c|A = a_1) - \mathbb{P}(C = c|A = a_0)).$$



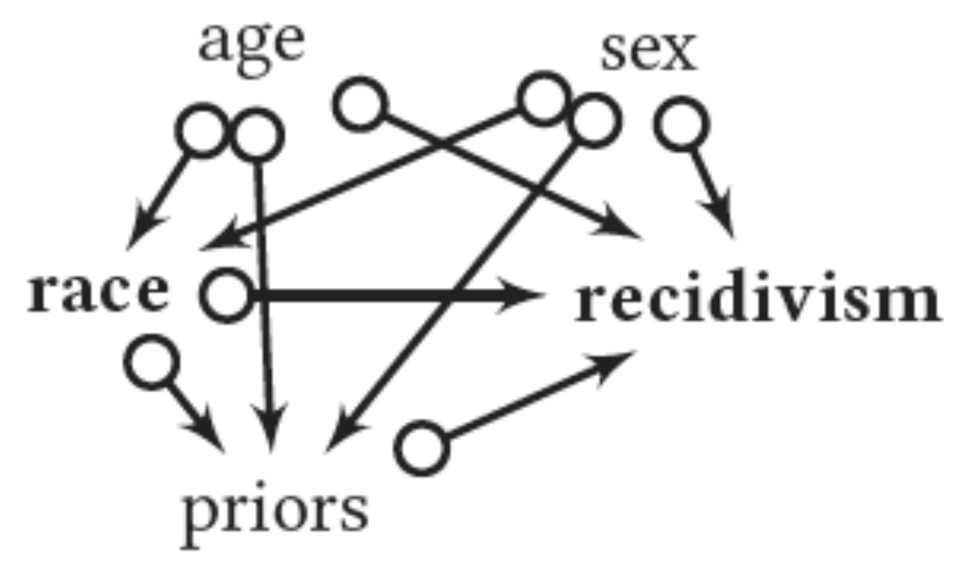
$$TE_{a_1,a_0}(y^+) = \mathbb{P}(y^+_{a_1}) - \mathbb{P}(y^+_{a_0})$$
$$= \sum_{c \in dom(C)} (\mathbb{P}(Y = y^+|A = a_1, C = c)$$
$$- \mathbb{P}(Y = y^+|A = a_0, C = c))\ \mathbb{P}(C = c)$$

$$NIE_{a_1,a_0}(y^+) = 0$$

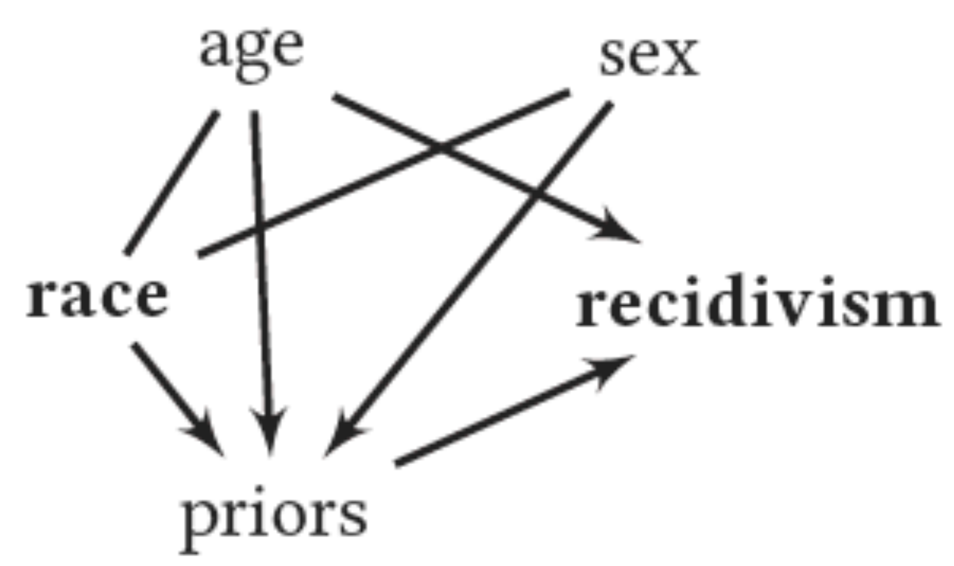
Causal Discovery for Fairness



(a) PC



(b) FCI



(c) GES

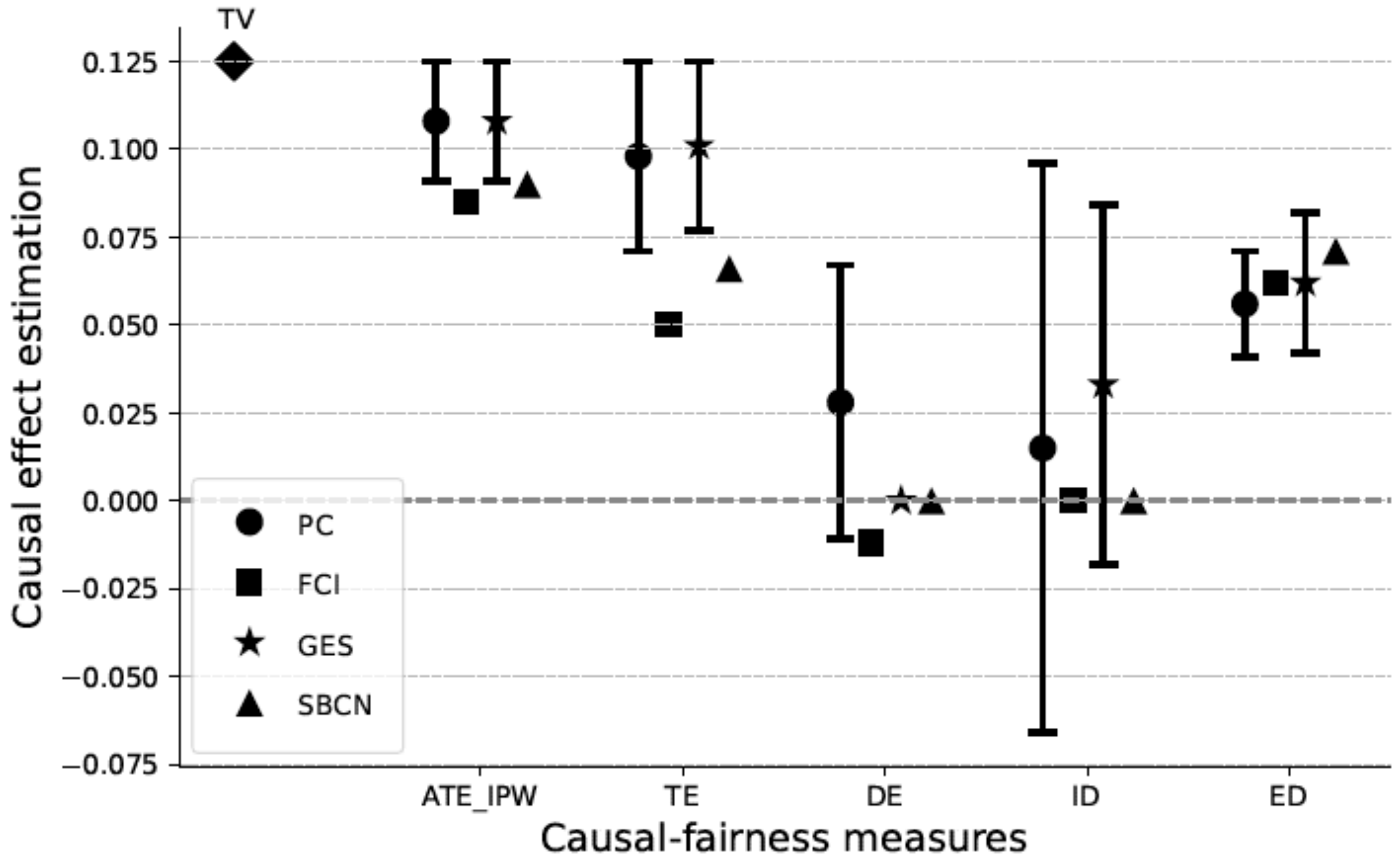
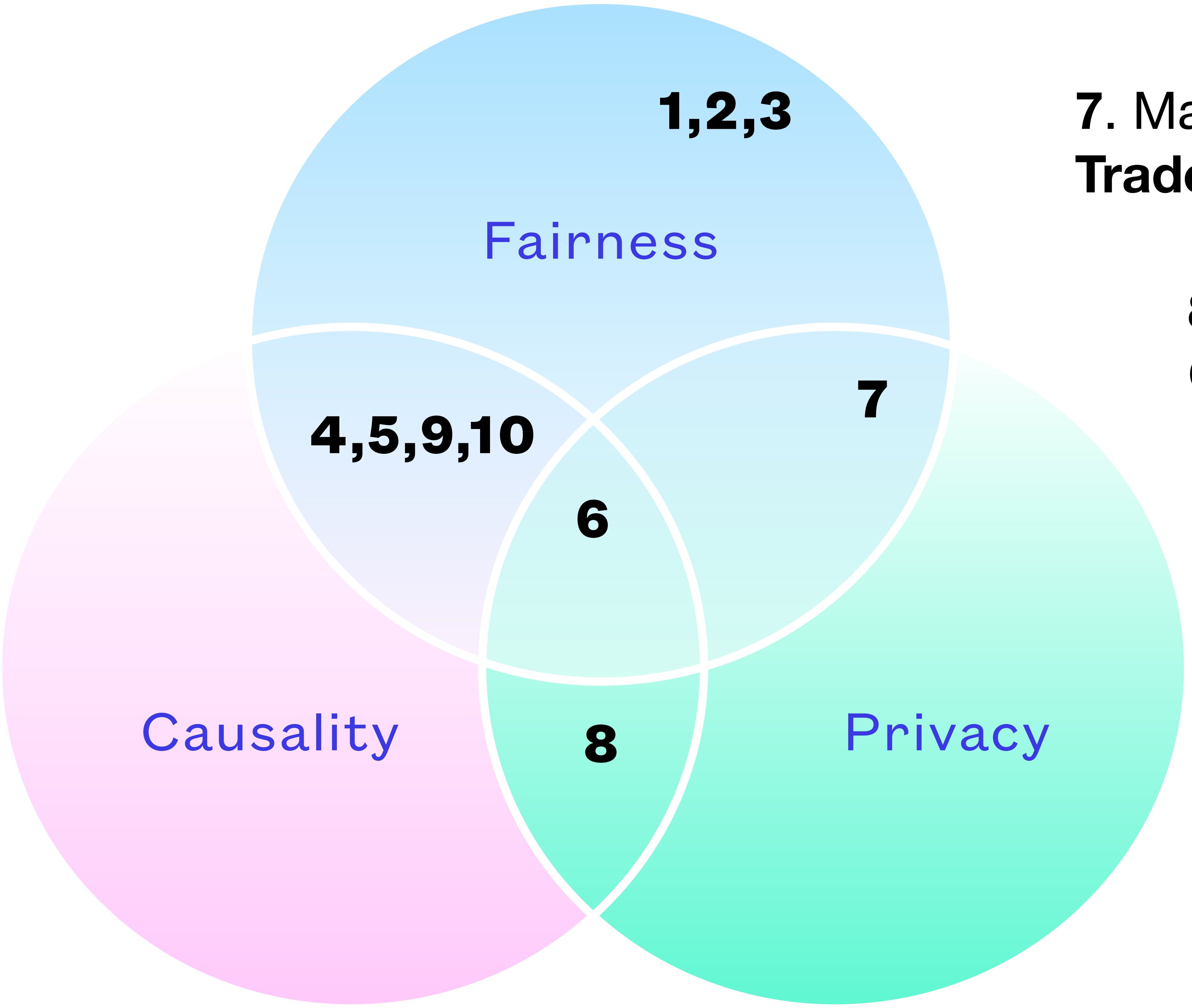


Figure 11: Estimation of causal effects of the Compas dataset based on PC, FCI, GES and SBCN.

In-progress



6. Binkytė, R., Palamidessi, C., Gorla, D.
BABE: Enhancing Fairness via Estimation of Latent Explaining Variables

7. Makhlouf, K., Arcolezi, HH., Palamidessi, C.
Trade-off between privacy and fairness

8. Binkytė, R., Arcolezi, HH, C., Zhioua, S., Palamidessi, C..
Causal Structure Preserving Local Differential Privacy

9. Binkytė, R., Makhlouf, K., Pinzón, Arcolezi, HH, C., Zhioua, S., & Palamidessi, C.
Designing a Causal Discovery Algorithm for Fairness

10. Zhioua, S., Binkytė, R.
Dissecting Machine Learning Bias with Causal Tools

Take-aways

- Causality is essential to reliably measure discrimination
- The two benefits of using causality in fairness:
    - Benefit 1: measuring discrimination accurately
    - Benefit 2: mediation analysis (distinguishing the different paths of discr.)
- Causality can be used to characterise sources of bias when measuring discrimination.

Thanks