

A purple sphere is positioned on a yellow triangular plane that is tilted upwards from left to right. The background is a solid pink color. The sphere and the plane are the central visual elements of the slide.

# The Role of Causality in Developing Fair AI Systems

A presentation by Maryam Badar

# Discrimination is Causal in nature...

- Discrimination can be causal in nature, meaning that it is often the result of systemic biases that are deeply ingrained in social, economic, and political structures.



Automatic Decision  
making system



Reason of  
Discrimination?



Historical data

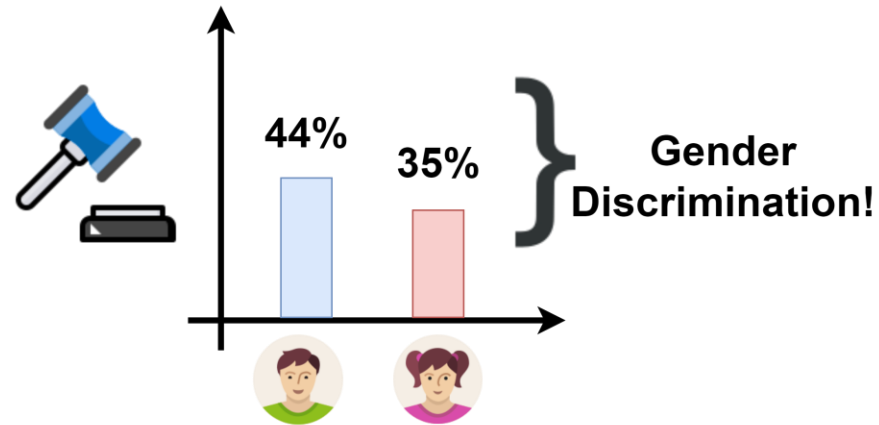
- By identifying the causal factors that contribute to discrimination, we can develop interventions and policies that address the root causes of the problem, rather than simply treating the symptoms.

# UC-Berkley—Simpson's paradox

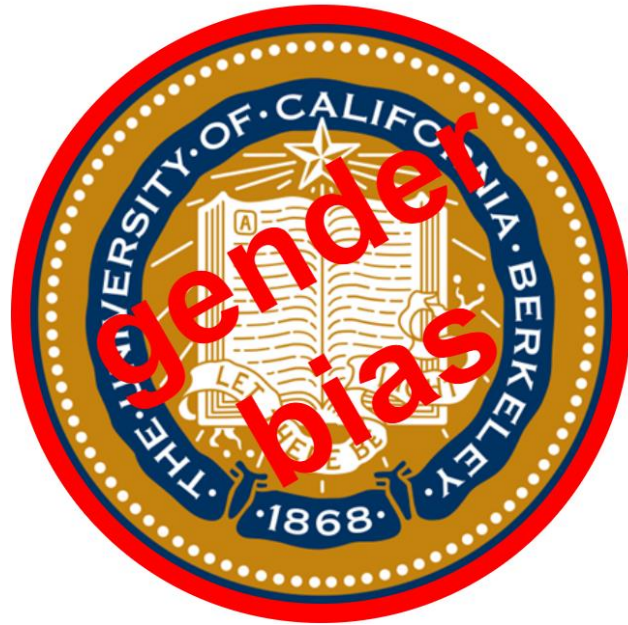
- The University of California, Berkeley in the 1970s feared a suspected gender bias in the outcomes of its graduate school admissions.



# UC-Berkley—Simpson's paradox





# UC-Berkley—Simpson's paradox





# UC-Berkley—Simpson's paradox




- Some departments had a higher proportion of male applicants, which made it more difficult for women to be admitted overall.
- Women tended to apply to departments that admitted a smaller percentage of applicants overall.
- Once the data was properly analyzed, it was found that there was no evidence of discrimination against women in the admission process at UC Berkeley.




# Hypothetical Example:

						
	no. of applications	no. of successful applications	success rate	no. of applications	no. of successful applications	success rate
Course P	80	38		20	14	
Course Q	20	2		80	16	
Total	100	40		100	30	

						
	no. of applications	no. of successful applications	success rate	no. of applications	no. of successful applications	success rate
Course P	80	38	47.5%	20	14	70%
Course Q	20	2	10%	80	16	20%
Total	100	40		100	30	



				 		
	no. of applications	no. of successful applications	success rate	no. of applications	no. of successful applications	success rate
Course P	80	38	47.5%	20	14	70%
Course Q	20	2	10%	80	16	20%
Total	100	40		100	30	

	<div>   </div>			<div>  </div>		
	no. of applications	no. of successful applications	success rate	no. of applications	no. of successful applications	success rate
Course P	80	38	47.5%	20	14	70%
Course Q	20	2	10%	80	16	20%
Total	100	40	40%	100	30	30%

# Conclusion

- This example highlights the importance of:
  - looking beyond overall aggregate data.
  - Carefully examining data by specific subgroups or categories, such as by department, to identify potential confounding variables and avoid drawing incorrect conclusions.
- In this case, simply looking at the overall admissions data would have led to the conclusion that there **was gender bias against women**, when in fact, the **bias was more complex** and varied by department.

# Causal Fairness notions: The Need

- Addressing root causes
- Mitigating unintended consequences
- Addressing confounding variables

# Preliminaries: Causal graph

- A causal graph is a directed acyclic graph where:
  - Nodes are the variables/attributes.
  - Edges denote direct causal effects.
  - Causal path is an acyclic sequence of adjacent nodes from the starting node to the last node.

# Preliminaries: Causal graph

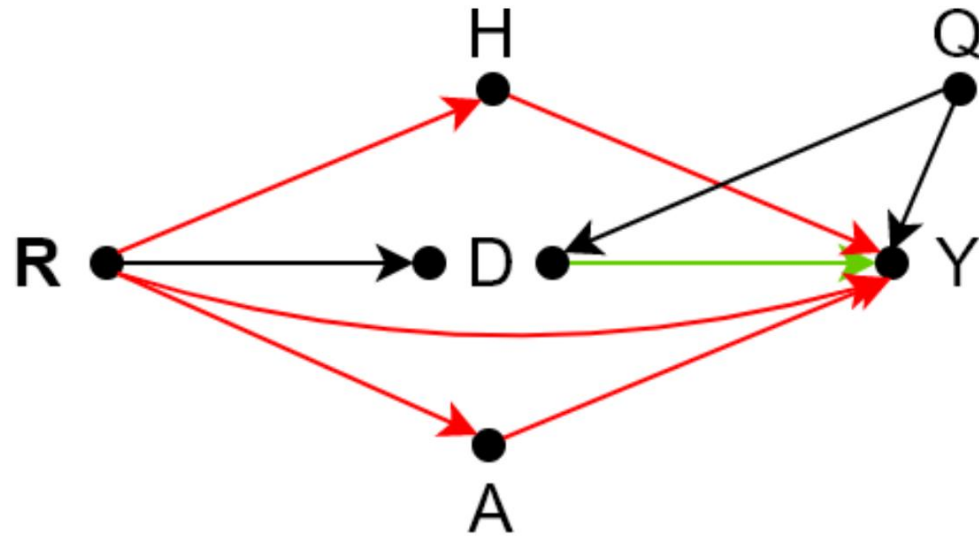
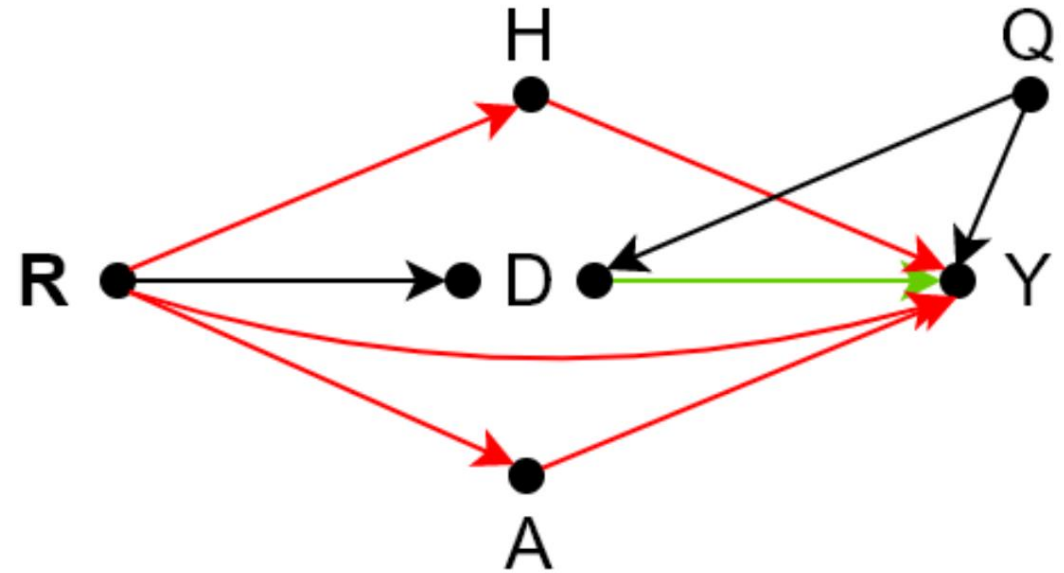


Fig: Causal graph of a university admission dataset representing fair and unfair causal pathways: R, D, Q, H, A, Y are causal variables representing “race”, “choice of department”, “qualification”, “hobbies”, “address”, and “admission outcome” respectively

# Preliminaries: Causal graph

- Direct effect:
  - $R \rightarrow Y$
- Indirect effect:
  - $R \rightarrow D \rightarrow Y$
  - $R \rightarrow H \rightarrow Y$
  - $R \rightarrow A \rightarrow Y$



# Causal fairness notions:

- Causal fairness notions can be segregated based on two criteria:
  - Counterfactual (CF)
  - Interventional (IF)



# Interventional (IF):

- IF measures fairness by quantifying the effect of sensitive attributes on the predicted outcome by intervening on the protected and non-protected attributes.
- IF approaches aim to relax some of the strong assumptions CF require to make them more practical in real-life settings.

# IF: Total Effect [Pearl2009]

- It is the causal version of the statistical parity group fairness notion.
- It measures the effect of changing sensitive attribute values on the outcome along all causal paths from sensitive attributes to the outcome.

$$TE_{p^+, p^-}(Y = y) = \mathbb{P}(y_{p^+}) - \mathbb{P}(y_{p^-})$$

# TE (Example)

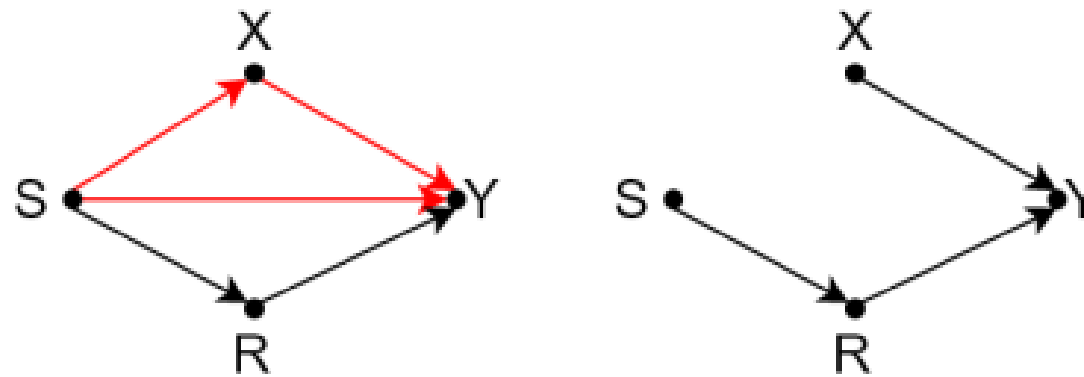
- In case of binary output, TE reduces to Average Treatment Effect (ATE)
- $TE = ATE = \frac{12}{24} - \frac{9}{24} = \frac{1}{8}$

**Table 1: A job hiring example with 24 applications.  $A$  is the gender (sensitive attribute) where  $A = 1$ : female,  $A = 0$ : male.  $C$  is the job type where  $C = 0$ : flexible time job,  $C = 1$ : non-flexible time job.  $Y$  is the hiring decision (outcome) where  $Y = 0$ : not-hired,  $Y = 1$ : hired.**

Female applicants (Treatment group)				Male applicants (Control Group)			
$i$	$A$	$C$	$Y$	$i$	$A$	$C$	$Y$
1:	1	0	1	13:	0	0	1
2:	1	0	1	14:	0	0	0
3:	1	0	0	15:	0	0	0
4:	1	0	0	16:	0	0	0
5:	1	0	0	17:	0	1	1
6:	1	0	0	18:	0	1	1
7:	1	0	0	19:	0	1	1
8:	1	0	0	20:	0	1	1
9:	1	1	1	21:	0	1	0
10:	1	1	1	22:	0	1	0
11:	1	1	1	23:	0	1	0
12:	1	1	0	24:	0	1	0

# IF: No unresolved discrimination [Kilbertus2017]

- It is a group fairness notion that focuses on the direct and indirect causal influence of sensitive attributes on the decision. It is satisfied when there is no direct path between the sensitive attributes and the outcome, except through a resolving/ admissible variable.



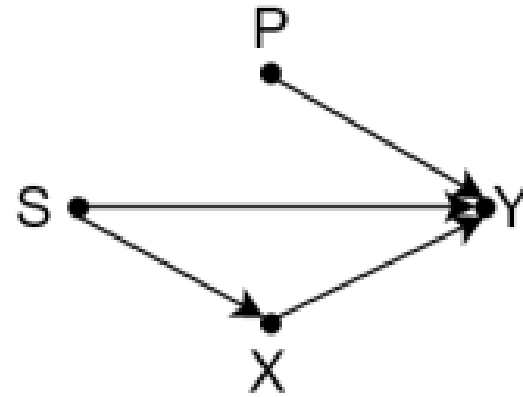
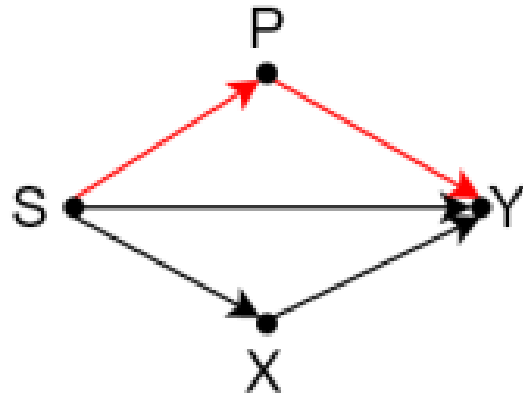
[Kilbertus2017]: Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. Advances in neural information processing systems, 30.

# IF: Proxy discrimination [Kilbertus2017]

- It is present in a causal graph when the path between the sensitive attributes and the outcome is intercepted by a proxy variable.
- A predictor  $Y$  avoids proxy discrimination if, for a proxy variable  $P$ , the following equation holds for all potential values of  $P$  ( $p_1, p_2$ ).

$$\mathbb{P}(Y = y | do(P = p_1)) = \mathbb{P}(Y = y | do(P = p_2)) \quad \forall p_1, p_2 \in dom(P).$$

# IF: Proxy discrimination



# IF: Individual direct discrimination [Zhang2016]

- This notion identifies direct discrimination at individual level.
- In any classification task, an individual is compared with  $n$  similar individuals sought from the protected group, denoted as  $D$  and  $n$  similar individuals from the non-protected group, denoted as  $D'$ . The similarity between individuals  $(i, i')$  is measured using causal inference.

$$d(i, i') = \sum_{k=1}^{|X|} |CE(x_k, x_{k'}) \cdot VD(x_k, x_{k'})|$$

$$CE(Y = y) = \mathbb{P}(Y = y | do(X)) - \mathbb{P}(Y = y | do(x'_k, X \setminus x_k)) \text{ and } VD(x_k, x_{k'}) = \frac{|x_k - x_{k'}|}{range}$$

# IF: Individual direct discrimination

- The individual in question is deemed not to be discriminated against if the difference between the positive prediction rates for the two groups ( $D$ ,  $D^-$ ) is under a predefined threshold.



# IF: Equality of Effort [Huan2020]

- This notion assesses discrimination by measuring the amount of effort required by the marginalized individual or group to reach a certain level of the outcome.

$$\psi_{G^+}(\gamma) = \psi_{G^-}(\gamma)$$

$$\psi_{G^+}(\gamma) = \operatorname{argmin}_{t \in T} \mathbb{E}[Y_{G^+}^t] \geq \gamma$$

- $G^+$  and  $G^-$  represent the set of individuals with  $S = p^+$  and  $S = p^-$  respectively which are similar to the target individual.

# IF: Interventional fairness [Salimi2019]

- A classification algorithm is interventionally K-fair if for any assignment of  $K = k$  and output  $Y = y$  the following equation holds:

$$\mathbb{P}(y_{p^+,k}) = \mathbb{P}(y_{p^-,k})$$

- K is a subset of attributes (V) except the sensitive attribute (S) and the outcome variable.

# IF: Justifiable fairness

- Justifiable fairness is a special case of interventional fairness, where we only consider those attributes for intervening that are admissible/resolving ( $E$ ) or a superset of admissible variables:

$$\mathbb{P}(y_{p^+,k}) = \mathbb{P}(y_{p^-,k}), k \supseteq E.$$

# Causal fairness notions:

- Causal fairness notions can be segregated based on two criteria:
  - Counterfactual (CF)
  - Interventional (IF)

# Counterfactual Criteria (CF):

- CF criteria assesses the effect of sensitive attributes on the predicted outcome by analyzing counterfactuals.
- If sensitive attribute (e.g. **S**: “gender”) is binary, it could take two values:
  - Protected (e.g. **p<sup>-</sup>** : female)
  - Non-protected (e.g. **p<sup>+</sup>** : male)

# CF: Counterfactual fairness [Kusner2017]

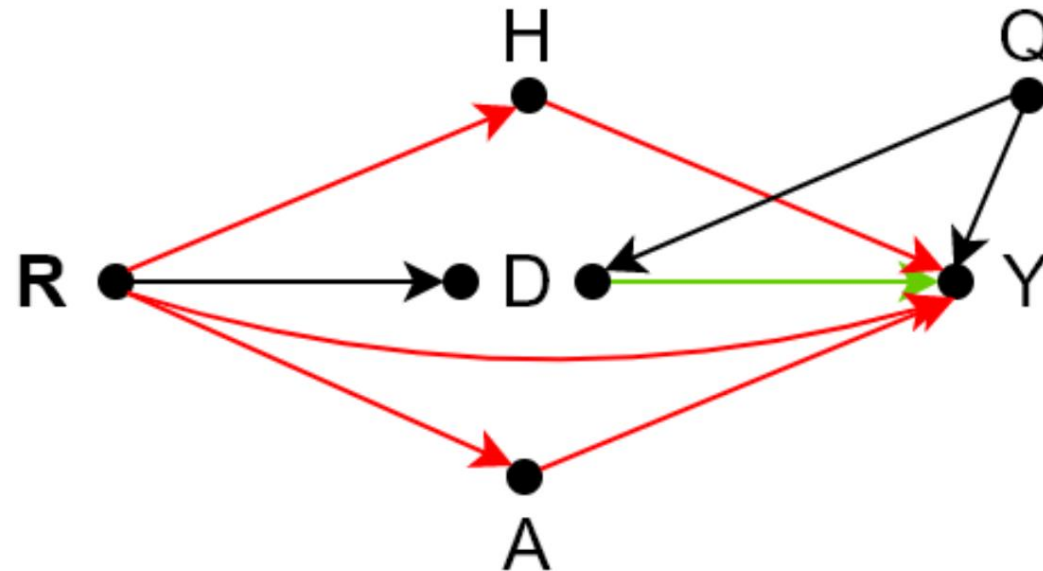
- This notion is achieved by a predictor  $Y$  for an individual if the probability of achieving the output  $Y = y$  remains the same if the value of sensitive attribute changes from  $p^-$  to  $p^+$ .

$$\mathbb{P}(y_{p^+} | X = x, S = p^+) = \mathbb{P}(y_{p^-} | X = x, S = p^-)$$

- This individual fairness notion assumes that the effect of sensitive attributes on the decision along all causal paths is unfair.

# CF: Counterfactual fairness

- The direct effect of race on the admission outcome is unfair.
- The indirect effect of race on the outcome through the “choice of department” variable is fair.



# CF: Path specific counterfactual fairness [Chiappa2018]

- This notion attempts to remove the causal effects of sensitive attributes on the outcome along only unfair causal paths.
- For a set of paths  $\lambda$ , path-specific counterfactual fairness exists if the following equation is satisfied.

$$\mathbb{P}(y_{p^+} | \lambda, p^- | \bar{\lambda}) = \mathbb{P}(y_{p^-})$$



# CF: PC-fairness [Wu2019]

- This is an additional path-specific counterfactual fairness notion for subgroups not just individuals. Given a set of paths ( $\lambda$ ) and a factual condition  $X = x$  ( $X \in V$ ), a predictor  $Y$  attains PC-fairness if it satisfies the following criteria:

$$\mathbb{P}(y_{p^+|\lambda, p^-|\bar{\lambda}}|X) = \mathbb{P}(y_{p^-}|X).$$

# CF: Counterfactual Equalized odds [Mishler2021]

- This fairness notion is satisfied by a predictor if the respective counterfactual false positive rates (cFPR) and counterfactual false negative rates (cFNR) of the protected group and non-protected group are equal which is not possible practically.
- $Diff^+(Diff^-)$  is the difference between cFPR (cFNR) of the protected and the non-protected group.

$$|Diff^+| \leq \varepsilon^+, \quad \varepsilon^+ \in [0, 1] \text{ and } |Diff^-| \leq \varepsilon^-, \quad \varepsilon^- \in [0, 1]$$

$$Diff^+ = cFPR(p^+) - cFPR(p^-) \text{ and } Diff^- = cFNR(p^+) - cFNR(p^-)$$

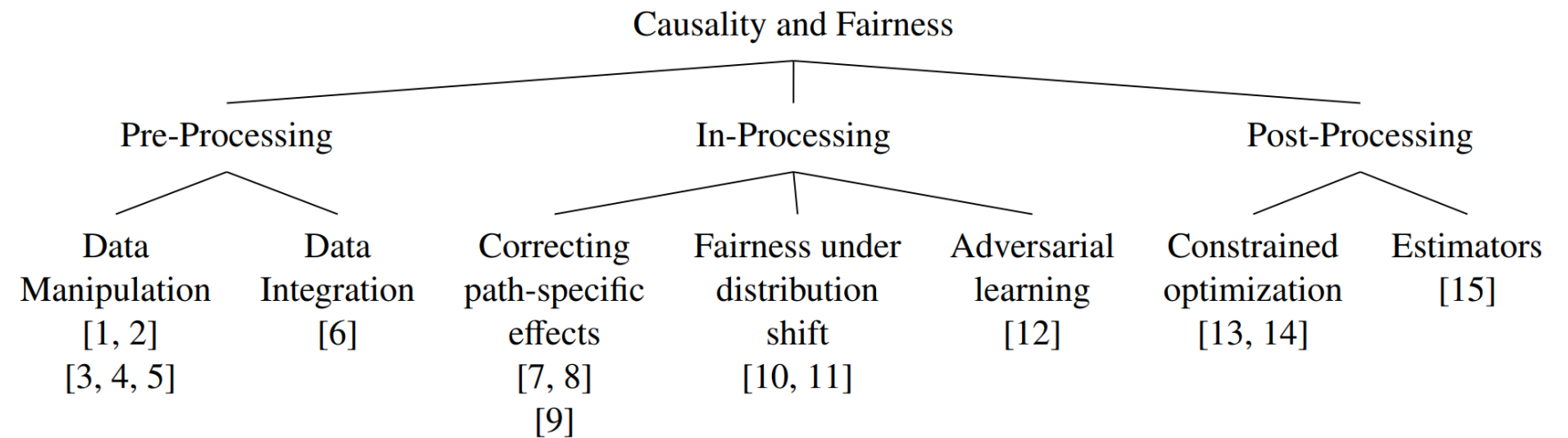
# CF: Causal Explanation formula [Zhang2018]

- This is a causal explanation method that helps in dividing the observed discrimination into three counterfactual effects: direct (DE), indirect (IE), and spurious effects (SE) of sensitive attributes on the outcome.

$$TV_{p^+, p^-}(Y = y) = |SE_{p^+, p^-}(Y = y) + IE_{p^+, p^-}(Y = y|S = p^-) - DE_{p^+, p^-}(Y = y|S = p^-)|$$

# Causality based fairness aware methods

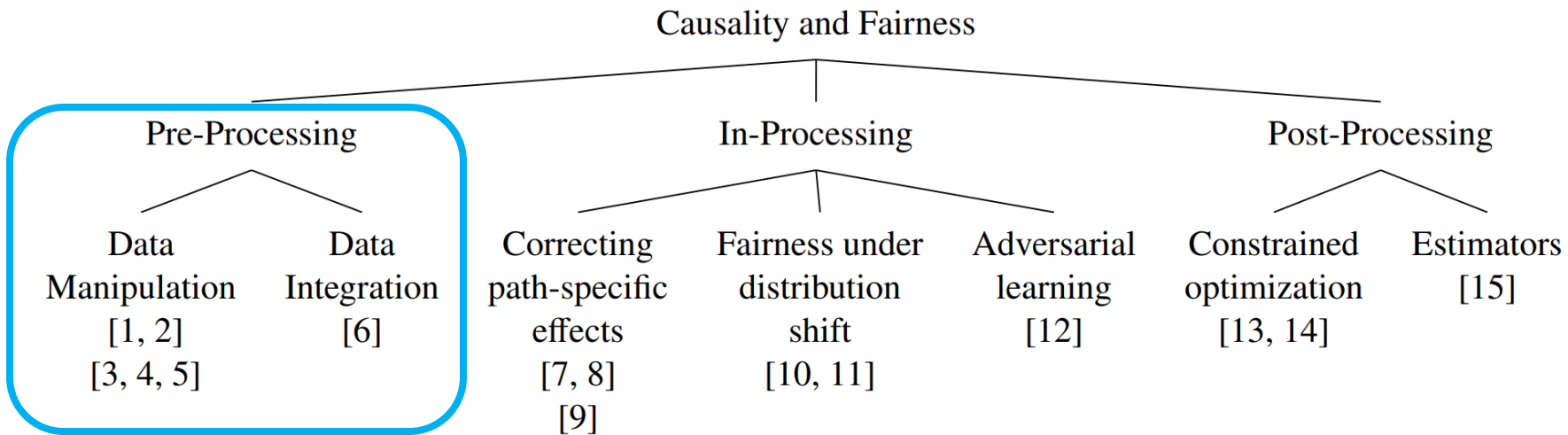
- Pre-processing
- In-processing
- Post-processing



**Figure 1: Segregation of causality-based methods for fairness-aware learning.**

# Pre-processing Methods

- These methods tailor the training data to make it bias free.



**Figure 1: Segregation of causality-based methods for fairness-aware learning.**

# Data Manipulation: Zhang et al. [1]

- Zhang et al. [1], calculate path-specific effects of sensitive attributes on the predicted outcome and compared them to a predefined threshold  $\tau$ .
- If the calculated path-specific effect exceeds  $\tau$ , this indicates the presence of direct and indirect discrimination.
- Later they eliminate both direct and indirect discrimination by generating a bias-free dataset through causal network manipulation that guarantee path-specific effects under  $\tau$ .

# Data Manipulation: Zhang et al. [2]

- The authors further extended their work to identify and handle the situations in which indirect discrimination cannot be measured because of the non-identifiability of certain path-specific effects.
- In such cases, the authors suggest setting an upper and lower bound on the effect of indirect discrimination.

# Data Manipulation: Zhang et al. [3]

- In this work, the authors detect direct and indirect system-level discrimination by measuring the path-specific causal effects of sensitive attributes on the outcome.
- To prevent discrimination, they modified the causal network to generate a new bias-free dataset.



# Data Manipulation: Xu et al. [4]

- The authors proposed a utility-preserving and fairness-aware causal generative adversarial network (CFGAN) to generate high-quality and bias-free data.

[4] D. Xu, Y. Wu, S. Yuan, L. Zhang, X. Wu, Achieving causal fairness through generative adversarial networks, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

# Data Manipulation: Salimi et al. [5]

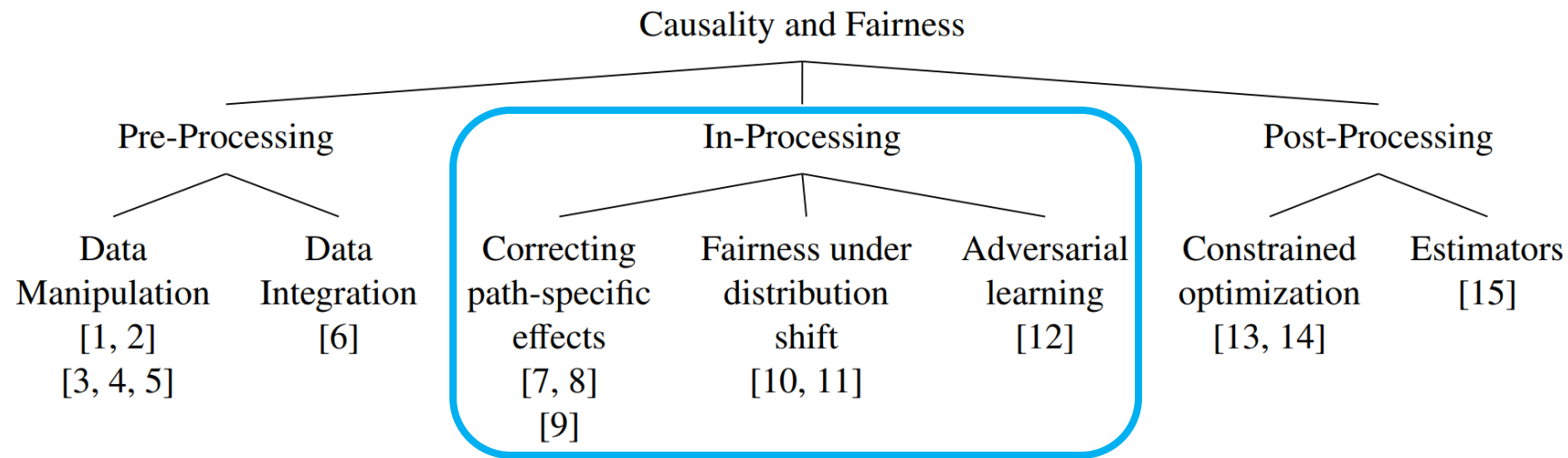
- They detected discrimination using interventional and justifiable fairness notions.
- To eliminate discrimination, they used causal dependencies between sensitive attributes and outcome variables to add and remove samples from the training data.

# Data Integration: Galhotra et al. [6]

- Data integration aims to combine data from various sources that capture a comprehensive context and enhance predictive ability.
- Galhotra et al. [6] modeled the problem of ensuring causal fairness in a learning task as a fair data integration problem.
- A conditional testing-based feature selection method was proposed that guarantees high predictive performance without adding bias to the dataset.

# In-processing Methods:

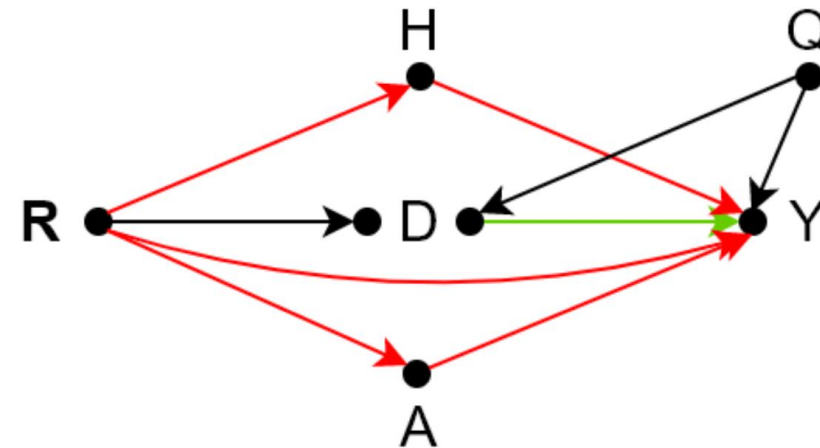
- These interventions adapt the learning algorithm itself to achieve fairness



**Figure 1: Segregation of causality-based methods for fairness-aware learning.**

# Correcting path-specific effects: Kilbertus et al. [7]

- The sensitive attributes can affect the outcome through both fair and unfair causal pathways.



- Fair causal pathway:  $R \rightarrow D \rightarrow Y$ .
- Unfair causal pathways:  $R \rightarrow H \rightarrow Y$  and  $R \rightarrow A \rightarrow Y$ .

# Correcting path-specific effects: Kilbertus et al. [7]

- Kilbertus et al. [7] used “proxy discrimination” and “unresolved discrimination” fairness notions to detect discrimination.
- They proposed to constrain the parameters of the learning algorithm so that the causal effects along both fair and unfair causal paths from sensitive attribute to outcome variable are removed.

# Correcting path-specific effects: Nabi et al. [8]

- The authors proposed to tackle the fair and unfair causal effect of sensitive attribute on outcome variable by constraining the path-specific effect during model training within a certain range.

# Correcting path-specific effects: Chiappa et al. [9]

- The authors presented a causal framework that ensures path-specific counterfactual fairness by **correcting the observations** of such **variables** that are **descendants of sensitive** attributes along only **unfair causal paths** so that only individual unfair information is eliminated while the individual fair information is retained.
- Hence improving predictive performance of the framework.



# Fairness under distribution shift: Singh et al. [10]

- The authors studied the problem of learning fair prediction models under covariate shift.
- They proposed a method based on feature selection such that the distribution of the sensitive attribute in the training dataset matches that in the testing dataset given the ground truth causal graph that explains the data.

# Fairness under distribution shift: Creager et al. [11]

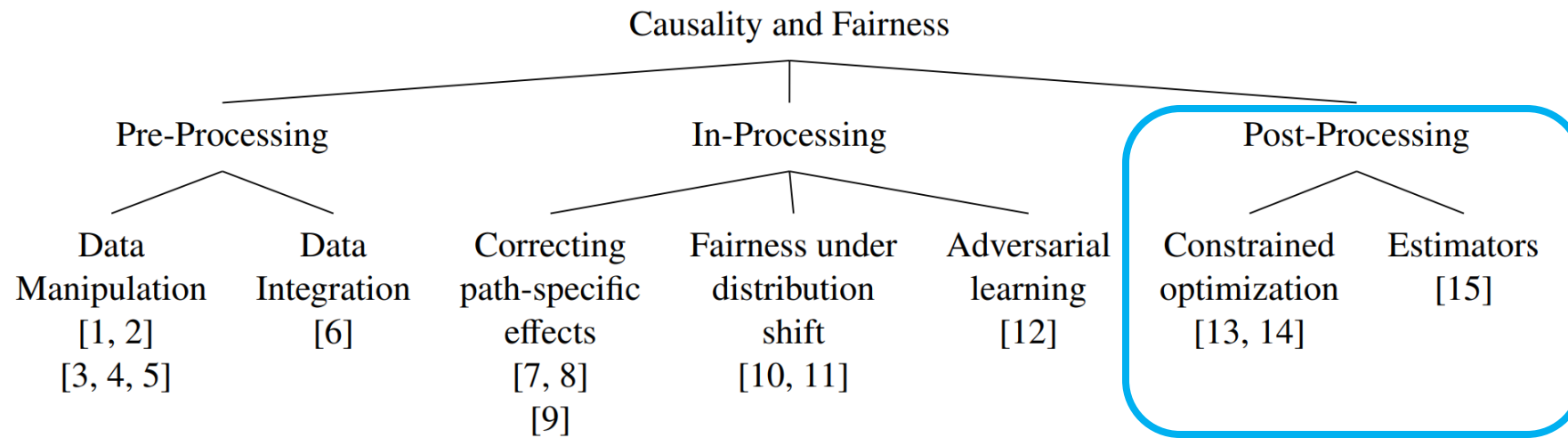
- Fairness concerns also surface when AI-based learners deal with dynamically fluctuating environments and produce long-term effects for both individuals and protected groups.
- Creager et al. [11] have proposed that in such dynamical fairness setting when the dynamic parameters are unknown, causal inference (causal Bayesian networks) can be utilized to estimate the dynamic parameters and improve off policy estimation from historical data.

# Adversarial Learning: Li et al. [12]

- The authors proposed an adversarial learning-based approach to achieve the goal of personalized counterfactual fairness for users in recommendation systems.
- They attempt to remove sensitive features information from the user embeddings by using a filter module and a discriminator module to make the learner's decisions independent of the sensitive features.

# Post-processing Methods:

- These methods tailor the outputs of the learner to achieve fair outcomes.



**Figure 1: Segregation of causality-based methods for fairness-aware learning.**

# Constrained optimization: Wu et al. [13]

- Wu et al. [13] proposed a method to bound the unidentifiability of counterfactual quantities and used c-component factorization to identify its source.
- They proposed a graphical criterion to determine the lower and upper bound on counterfactual fairness in unidentifiable scenarios.
- Finally, they proposed a post-processing method to reconstruct the decision model to achieve counterfactual fairness.

# Constrained optimization: Kusner et al. [14]

- The paper proposed a framework for identifying discriminatory impacts in decision-making processes. The framework takes into account the different **stages** of the decision-making process, from **data collection** to the **implementation of the decision**, and **identifies the potential sources of discrimination** at each stage.
- The authors achieved counterfactual fairness by constraining the beneficial effects obtained by an individual under a limit depending on the sensitive attribute of the individual.

# Doubly robust estimators: Mishler et al [15]

- Authors proposed a post-processed predictor, estimated using doubly robust estimators, to achieve the counterfactual equalized odds fairness notion.
- Through experiments, they also proved that their method has favorable convergence properties.

[15] A. Mishler, E. H. Kennedy, A. Chouldechova, Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds, in: FAccT, 2021, pp. 386–400.

# Conclusion and future work

- A unique trait of fairness research is the usage of multiple metrics to define/measure it.
- Choosing the most appropriate notion of fairness applicable to a particular situation is an important task.
- Even if a fairness notion is found suitable for a scenario, it may not be applicable due to the problem of identifiability, as Pearl's SCM framework requires causal quantities, counterfactuals, and interventions, to be identifiable.



# Conclusion and future work

- The pre-processing methods discussed here are only applicable in a static environment where all data are available in advance.
- An interesting future direction could be the extension of such methods for online learning, where not all data are available beforehand.
- Most of the causality-based fairness solutions discussed above rely on the assumption that the underlying data is independent and identically distributed (IID).
- However, real-world use cases include non-IID data, therefore, another future direction could be to design causality-based decision support systems which relax the assumption of non-IID data and provide non-discriminatory predictions.

# Conclusion and future work

- All research done in the field of causal fairness is connected to classification tasks.
- Little to no research has been done to achieve causal fairness in community detection, word embedding, named entity recognition, representation learning, semantic role labeling, language models, and machine translation.



Thank you for your attention