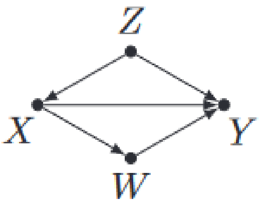
Fairness in Decision-Making—The Causal Explanation Formula

- Can decompose discrimination into (at least) two parts
    - *Direct*: explicit use of protected attribute to perform discrimination
        - E.g. voting rights, unequal payment
    - *Indirect*: usage of a proxy to perform discrimination
        - E.g. redlining (zip code as proxy for race)
- Consider these types of discrimination while addressing various fairness metrics that have been proposed
    - "Despite all the recent progress in the field, there is still not a clear understanding of the various metrics used to evaluate each type of discrimination individually. In practice, this translates into the current state of affairs where the fairness criterion is, almost invariably, chosen without much discussion or justification."
- Build on Pearl's causal framework
    - Just as a causal effect can be decomposed into direct, indirect, and spurious effects, so can discrimination
    - Specifically denote these as *counterfactual direct (Ctf-DE), indirect (Ctf-IE),* and *spurious (Ctf-SE) effects*
    - Provide a *causal explanation formula* that combines these effects into an overall measure of discrimination, which in turn can be separated into each component
    - Use tools to analyze differences in procedural vs outcome fairness
- Overview of structural causal model setup…(DAG, natural direct effect, natural indirect effect, etc.)
- Introduction of counterfactual effects of discrimination
    - Religious discrimination hiring example
        - *X*: religious belief (protected attribute)
        - *Z*: educational background (confounder)
        - *W*: proximity to places of worship (mediator)
        - *Y*: hiring decision (outcome)
        - 

(a)
    - If a company hires based on *Z* and is nevertheless accused of discrimination on *X* or *W*, existing measures can only check for direct and indirect discrimination but not for spurious discrimination or a combination of metrics
        - Motivates deeper analysis of disparities

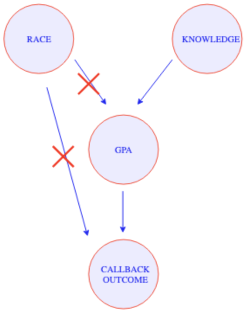
Disparate Causes Pt. I

Lill Hu

I. Algorithmic Fairness
   - A. We have to ask: what are the cause-effect relationships between relevant attributes that lead to a decision? More bluntly: ***does the attribute of race cause a particular decision outcome?***

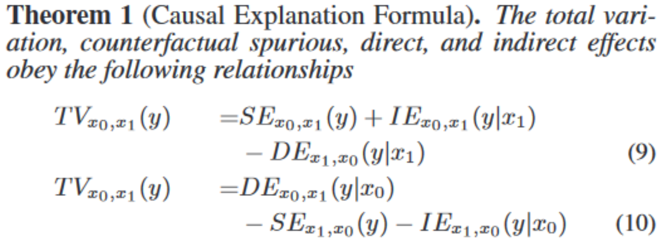
II. The Direct Effect View
   - *A. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination"*
   - B. "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race... and everything else had been the same."
   - C. Can everything else stay the same?
     1. rests upon an incorrect theory of what race is
     2. This view, I argue, reduces wrongful discrimination to something like irrationality in decision-making.
     3. a social constructivist understanding of race

III. The Path Specific Effect View
   - A. As I see it, they come to us as normative valuations about why certain observational data of decision outcomes are wrongfully racially biased. These valuations might be explained in causal terms, but they are not founded on causal reasoning. Normative assessments of observational outcomes inform the structure of our causal diagrams, not the other way around.



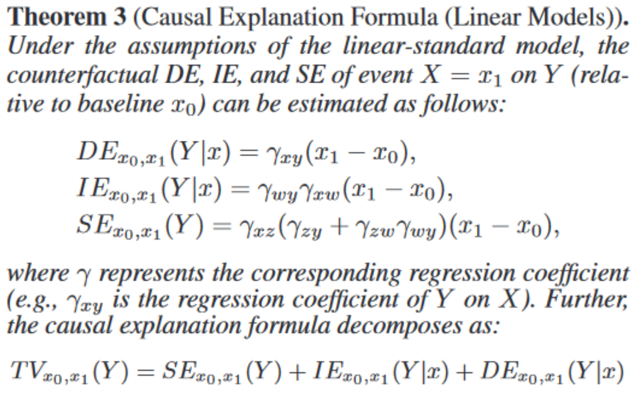
Kohler-Hausmann: *"The ideal experiment to detect discrimination in the counterfactual causal model is one in which the researcher.... [zeroes] out the average differences in relevant variables that were produced by the real lived institutions of racial orders."*

Thinking causally about the effects of race can help inform and elucidate some of our normative judgments about what we ought to do in our efforts toward racial equality, and the specific task of drawing a causal diagram might even be productive for our public discourse. But if the goal of our causal fairness methods is to build decision procedures that are racially just, we should start by asking what type of outcomes we would expect a racially just procedure to yield.

  - **Counterfactual Direct Effect**: given *X*=*x* in both cases, the difference in probabilities when fixing x_1 and setting the mediator value *W* to that obtained under x_0 versus fixing everything under x_0
    - $$DE_{x_0,x_1}(y|x) = P(y_{x_1,W_{x_0}}|x) - P(y_{x_0}|x)$$
    - Can also think of this as holding the mediator *W* constant and examining the effect of *X* on *Y* (as summarized by the causal path X -> Y)
  - **Counterfactual Indirect Effect**: given *X*=*x* in both cases, the difference in probabilities when fixing x_0 and setting the mediator value *W* to that obtained under x_1 versus fixing everything under x_0
    - $$IE_{x_0,x_1}(y|x) = P(y_{x_0,W_{x_1}}|x) - P(y_{x_0}|x)$$
    - Can also think of this as holding the protected attribute *X* constant and changing the mediator *W* to examine the effect of *W* on *Y* (to measure effects of the causal path X -> Z -> Y)
  - **Counterfactual Spurious Effect**: the difference in probabilities when fixing x_0 among individuals with x_1 versus outcomes among individuals with x_0
    - "measures the difference in the outcome *Y* = *y* had X been x_0…for the individuals that would naturally choose *X* to be x_0 versus x_1"
    - Tests for: back-door paths
    - Intuition: if counterfactually modifying the protected attribute still results in a difference between the two groups, then there must be some confounder *Z* that accounts for these differences
- Causal Explanation Formula
  - Non-parametric formula
    - 

      Theorem 1 (Causal Explanation Formula). *The total variation, counterfactual spurious, direct, and indirect effects obey the following relationships*

      $$TV_{x_0,x_1}(y) = SE_{x_0,x_1}(y) + IE_{x_0,x_1}(y|x_1) - DE_{x_1,x_0}(y|x_1) \quad (9)$$

      $$TV_{x_0,x_1}(y) = DE_{x_0,x_1}(y|x_0) - SE_{x_1,x_0}(y) - IE_{x_1,x_0}(y|x_0) \quad (10)$$
  - More details on how to identify and estimate from data…
  - Parametric formula
    - 

      Theorem 3 (Causal Explanation Formula (Linear Models)). *Under the assumptions of the linear-standard model, the counterfactual DE, IE, and SE of event $X = x_1$ on Y (relative to baseline x_0) can be estimated as follows:*

      $$DE_{x_0,x_1}(Y|x) = \gamma_{xy}(x_1 - x_0),$$
      $$IE_{x_0,x_1}(Y|x) = \gamma_{wy}\gamma_{xw}(x_1 - x_0),$$
      $$SE_{x_0,x_1}(Y) = \gamma_{xz}(\gamma_{zy} + \gamma_{zw}\gamma_{wy})(x_1 - x_0),$$

      where γ represents the corresponding regression coefficient (e.g., γ_{xy} is the regression coefficient of Y on X). Further, the causal explanation formula decomposes as:

      $$TV_{x_0,x_1}(Y) = SE_{x_0,x_1}(Y) + IE_{x_0,x_1}(Y|x) + DE_{x_0,x_1}(Y|x)$$

    - Also consider case with unobserved confounding
- Experiments
    - Illustrate how to use decomposition formula to identify different types of discrimination from data
    - Also shows how to design reparatory policies by responding to direct effect or all effects of discrimination
        - Policies need to reside within a feasible region such that the effects of the policy reside within an amount of discrimination up to the original amount being addressed
        - “Narrow tailoring”
- Afterword
    - Other causal inference work in this area
        - Kusner et al. (2017) “Counterfactual Fairness”
        - Kilbertus et al. (2017) “Avoiding Discrimination through Causal Reasoning”
        - Nabi and Shpitser (2018) “Fair Inference on Outcomes”
        - Chiappa (2019) “Path-Specific Counterfactual Fairness”
        - Coston et al. (2020) “Counterfactual Risk Assessments, Evaluation, and Fairness”
    - Drawbacks/Open Questions
        - What to do in case of unobserved confounding?
        - Does isolating discrimination/unfairness to a certain path make it ok?
            - E.g. Nabi and Shpitser explore the COMPAS problem and note the following: “We are interested in predicting whether a defendant would reoffend using the COMPAS data. For illustration, we assume the use of prior convictions, possibly influenced by race, is fair for determining recidivism. Thus, we defined discrimination as effect along the direct path from race to the recidivism prediction outcome.”
        - The definitions introduced by Zhang and Bareinboim (and others in this area) assume it is possible to counterfactually modify protected attributes while leaving all else constant (such as covariates and/or mediators)
            - Is this (always) possible to do? Does it matter?
            - Perhaps a good segue into Hu and Kohler-Hausmann’s works
            - Also may want to check out “What’s Sex Got to Do With Fair Machine Learning?” by Hu and Kohler-Hausmann

HANDOUT – May 5, 2023

Handout: "Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination"

CONNY KNIELING
Email: cok22@pitt.edu

OVERVIEW

The common model of discrimination—the counterfactual causal model—is wrong and should be replaced by her proposal.

THE COUNTERFACTUAL CAUSAL MODEL

- Common model of discrimination, in sociology and law
- *"Discrimination, on this account is detected by measuring the "treatment effect of race," where treatment is conceptualized as manipulating the raced status of otherwise identical units (e.g., a person, a neighborhood, a school). Discrimination is present when an adverse outcome occurs in the world in which a unit is "treated" by being raced—for example, black—and not in the world in which the otherwise identical unit is "treated" by being, for example, raced white"* (p. 1167)
- Detecting discrimination as the treatment effect of race in the counterfactual sense
- Notion of Causality at work: detecting causation via manipulation
- Race not a manipulative variable on the unit, but a trait of the unit of interest, race as cause-qua-treatment

ISSUES WITH THE COUNTERFACTUAL CAUSAL MODEL

- **Old: Methodological Issues**
  (a) Issues of Manipulability (are by now somewhat addressed)
  (b) Issues of Temporality (are by now somewhat addressed)
  Still: talking about race as (presupposing race to be) an attribute or trait

- **New: Sociological and Normative Objection (theoretical, not empirical or practical)**
  - Incompatible with the constructivist theory of race

  Her objections: based on/presupposes a flawed theory of race; therefore misunderstands/misconceives
  (a) what the concept race references, how it produces effects in the world,
  (b) what we mean when we say it is bad to make important decisions "based on race"
  ("cements an already predominant and problematic understanding of race in public and legal discourse: one that is distressingly dehistoricized and desocialized")
  There is no non-backtracking way to specify "treatment of race"

  - entire logic of the CCM is to define discrimination as a form of irrationality by stripping away the system or any constitutive explanations

  Instead: major conceptual shift

AUDIT STUDIES

- proceed as if the aim of the exercise is to get a pure treatment effect of race presuming there is an objective trait to be gotten at

Objections:
(a) do not measure objectively bounded treatment effects of race and race alone
(b) results are recognizable as discrimination because of new account
(c) experimental methods are not the "gold standard" for detecting discrimination

HER PROPOSAL

An adequate theory of discrimination must rest upon (two tools she proposes):

(1) Thick ethical concepts (Bernard Williams): require complex, social knowledge to use or decode them
Terms that describe and evaluate
"To morally evaluate an action with a thick ethical concept communicates information about the way in which the action is bad that relies on institutional and cultural facts" (p. 1171)

(2) Constitutive explanations: (ground thick ethical concepts)
"A constitutive claim accounts for the capacities of complex systems by reference to their constitutive elements: the parts and organization that make the system what it is." (p.1172)
- Proffer counterfactual dependence and explanations

"Combining these two conceptual components yields a definition of discrimination as an action or practice that acts on or reproduces an aspect of the category in a way that is morally objectionable. It is a thick ethical concept that—to express the distinctive wrongfulness of the action vis-à-vis the category—must rest upon an account of the system of social meanings or practices that constitute the categories at issue." (p.1172)

- Entails empirical and normative elements which are black boxed in this article