# Discrimination Through the Lens of Philosophy of Science

Naftali Weinberger

March 5, 2024

**Abstract**

In Winter 2023-4, I taught a Masters seminar called "Discrimination Through the Lens of Philosophy of Science" at the Munich Center for Mathematical Philosophy, LMU Munich. The course covered topics related to discrimination that intersected with more general themes from philosophy of science and involved a wide range of readings from philosophy, sociology, causal inference, legal studies, and the philosophy of artificial intelligence. I was surprised by the extent to which these readings from independent disciplines and sub-disciplines engaged with a common set of questions and debates. Because there does not appear to be a textbook or course that covers the range of topics we considered, I here give a summary of the course in the hope of encouraging others to develop such courses and textbooks. Additionally, I think it would be possible to develop an introductory philosophy of science course called "Philosophy of Science Through the Lens of Discrimination" but such a course would only be feasible given a textbook or alternate readings, as the syllabus for this course was ambitious even for an MA seminar.

## 1 Appearance Discrimination

In the first class we discussed short story "Liking What You See: A Documentary" by Ted Chiang. The story imagines a simple and reversible procedure that allows individuals to no longer perceive beauty and physical attractiveness, and reports on a debate surrounding a vote on whether a university should make this procedure mandatory for all attending students. The goal of this first class was to get the students thinking about whether appearance discrimination would count at discrimination under these circumstances, and more generally, about what determines which social categories call for legal protections against discrimination. We distinguished between more narrow notions of discrimination that focus on **socially salient categories** such as race, gender, and religion, and broader notions that focus on unfairness, arbitrariness, or inequality. The story also provided an opportunity to think about how norms of discrimination can be sensitive to what is technologically or sociologically possible, and how the legal question of which types of discrimination should be regulated can come apart from questions about the morality of choices made voluntarily by individuals. Finally, I was pleased that the students raised the question about whether it was coherent to talk about disentangling status-oriented traits from other social categories, as the issue of whether it is possible to **disentangle** race and gender from their social manifestations would become a dominant theme in the course.

In principle, one could spend many weeks discussing the questions of what makes discrimination distinctively wrong, and of whether discrimination applies differently to socially salient categories. These questions were always in the background of our discussion, but in my version of the course we mainly focused on racial discrimination, with some briefer discussions of gender, sexual orientation, and disability.

## 2 Social Structural Explanation

Explanation is one of the core topics covered in the last century of philosophy of science. Discrimination provides an opportunity to focus on explanation in the social sciences and on methodological individualism. This class focused on a debate between Sally Haslanger (2016) and Lauren Ross (2023) on social structural explanation. We first had a discussion about why it matters whether an explanation is structural in order to get a sense of what an account of such explanations should seek to do. An explanation's being structural matters both for the question of whether individuals are responsible

for an outcome as well as how negative outcomes should be addressed. To the extent that a problem is structural, solutions to the problem require collective action and an understanding of wider historical and sociological mechanisms, and a myopic focus on individual responsibility is potentially both ineffective and victim blaming. We then turned to the debate. Haslanger gives a heterogeneous range of examples that purportedly involve structural explanation and presents an account on which such explanations should be understood in terms of part/whole relations. Ross, in contrast, argues that structural explanations can be understood causally, focusing on examples in which an individual faces constraints. For example, an individual who is unable to get to work as a result of there being no available bus. From a causal modeling perspective, Ross' account treats constraints as **interactive causes** – causes whose influence on an effect depends on the presence of other causes of that effect. In Ross's case, an individual's desire to go to work can only lead to success if the relevant enabling factors are in place, such as there being a bus route.

This section served as the first example of the trade off between precision and adequacy involved modeling complex social phenomena. Ross' account is formally more rigorous, but only focuses on particular types of cases. Her account seems especially ill equipped for dealing with structural explanations that appeal to social norms. On the other hand, the diversity of examples in Haslanger's account is not necessarily a virtue, since the set is so heterogeneous it is unclear that explicating them using the part-whole analogy does much work. Since the part-whole relationship is an example of constitutive relevance, in which the whole is constituted by its parts, this debate presages that between **causal and constitutive approaches** to discrimination, which would persist throughout the course.

The Ross article was also the first time we encountered **interventionism** and **causal modeling**. A curious feature of her discussion is that while causes have different properties from constraints (e.g. causes are shorter-term and constraints are longer-term), within a causal model causal interaction is symmetric (X interacts with Z in causing Y iff Z interacts with X in causing Y). This is not an inconsistency, but simply means that any distinction between causes and constraints is not internal to the model, but must be specified by conditions stipulated externally to it. It would nevertheless be illuminating to see whether the features differentiating causes and constraints could themselves be incorporated within a causal model.

# 3 Social Constructivism

## 3.1 Week 1: Social Ontology

In discussions of racial discrimination, one often encounters the worry that talk of the causal effects of race commits one to race being problematically real or even biologically essential. From a philosophy of science perspective, this is strange for two reasons. First, a widely employed criterion for thinking about whether a property is real is whether it figures in non-accidental generalizations, and it seems plausible that race could figure in such social generalizations without being biologically real. Second, to the extent that the dominant view of race is that it is **socially constructed**, social constructivism is typically a realist theory. That is, it tries to explain what race is rather than to argue for its elimination as a scientifically legitimate concept. The main readings for this week were Michael Root's "How we divide the world" (2000) and Ron Mallon's "Passing, Travelling and Reality" (2004). The aim of the former piece was to focus on the issue about scientific generalizations regarding race, and the aim of the latter was to give students background regarding the nuances of social constructivism.

The Mallon piece has many moving parts and thus requires a great deal of attention to see how it all fits together. Nevertheless, it yields significant dividends by drawing distinctions that would continue to be be relevant throughout the course. The basic strategy of the paper is to present three desiderata for a social constructivist account of race and to argue that no single account can satisfy all three desiderata. One desideratum is the ability to account for the phenomenon of passing – successfully presenting oneself as being of a different race than one actually is. Mallon argues that "experiential" accounts that identify race with its social effects cannot account for passing, since successfully passing entails inducing the same social effects as if one were of a certain race, without actually *being* of that race. In contrast, "folk objectivist"/"thin race" constructivist accounts enable one to distinguish between the **central criteria** that socially determine one's race, and **indicative criteria** that are not essential to race, but which enable members of a society to infer an individual's race. Passing occurs when one's indicative criteria don't match one's core criteria. Yet such account do not satisfy the

desideratum that "race does not travel", which claims that racial categorization only apply relative to a cultural context (e.g. it is nonsensical to ask whether Cleopatra was African American). The reason is that the type of criteria employed will assign a race individuals in other cultures as well (which is a different question from whether the people in those cultures would self-identify as that race). Mallon then considers a third, "institutional" constructivist account that can satisfy the first two desiderata, but which does not satisfy the third desideratum, which is that race be real. He concludes that no account can satisfy all three desiderata, but suggests that it is not a problem as long as we are clear about which constructivist race concepts are being used in which contexts.

As you can see, this all gets very complicated. But the distinction between experiential and folk objectivist accounts nicely sets things up for later discussions of causal versus constitutive accounts of racial discrimination. Experiential views are constitutive views, since they do not view race as causing discriminatory experiences, but rather as being constituted (in part) by those experiences. This is why race does not travel on such accounts; in different cultures racial categories are constituted by different and incommensurable packages of experiences. Folk objectivist accounts reject such holism by differentiating between central and indicative criteria and allowing the effects of race to be causally mediated via the indicative criteria.

Although it was only a supplemental source, the This American Life segment "Occam's Razor" (https://www.thisamericanlife.org/214/family-physics/act-one-9) was useful for contemplating the experiential account of race. It describes the journey of a man with stereotypically black features who was raised to believe he was the biological son of two white parents.

## 3.2 Week 2: Methods and Metaphysics

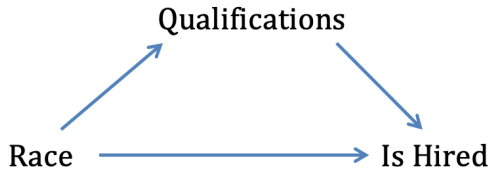
The aim of this week was to begin connecting issues of ontology to empirical methods for studying discrimination. Khalifa and Lauer (2021) provided a useful starting point for this discussion. A common argument pattern supporting scientific realism claims that the existence of certain entities provides the best explanation for why scientific theories positing those entities are successful. Khalifa and Lauer consider and reject a version of this argument, which claims that a realist constructivist theory best explains the role of race in successful social sciences. The paper is useful for situating racial ontology in the context of more general philosophy of science discussions and, more specifically, for viewing a connection between what race is and how it is empirically studied. Additionally they provide several arguments against constitutive accounts of race.

We then turned to Sen and Wasow's influential article "Race as a Bundle of Sticks" (2016). This was useful for introducing both key experimental designs for testing discrimination, such as **audit studies** and the implicit association test as well as the general issues surrounding the causal interpretation of race. Sen and Wasow take the interpretation of race as being socially constructed as having a direct implication for its **experimental manipulability**, which is often taken as a criterion for its being a cause. Specifically, they claim that although race as a whole cannot be manipulated it has parts that can be ("sticks" in the race bundle). The sticks/bundle metaphor points to a constitutive relevance relation, although we discussed in class how the authors use two distinct metaphors for understanding different types of experiments. Whereas for exposure experiments such as audit studies, the sticks are described as "proxies" for the bundle, for within-group experiments the sticks are described as "explaining" the bundle. The versatility of the stick/bundle metaphor might also be its weakness, as one might wonder how much it illuminates about the heterogeneous experiments to which it is applied. Additionally, we noted how the bundle metaphor does not lend itself naturally to Mallon's distinction between indicative and core properties.

In the next part of the course we switched from ontology to methodology. Nevertheless, the question of what social constructivism implies about whether discrimination ought to be analyzed causally recurred throughout the course.

# 4 The Causal Status of Race

## 4.1 Week 1: Against Causal Counterfactual Thinking About Discrimination




Issa Kohler-Hausmann's "Eddie Murphy and the Dangers of Causal Thinking about Detecting Racial Discrimination" (2018) moves past simply talking about whether race is a cause and towards thinking about whether a causal approach really makes sense given sociological and legal theories of discrimination. The article has lots of moving parts, and, to be honest, I didn't fully appreciate its core point until after a fruitful conversation with Kohler-Hausmann. To better understand the piece, it helps to consider the causal graph above, which represents an audit study. I take Kohler-Hausmann's most penetrating criticisms of causal interpretations of discrimination to derive from concerns as to whether (perception of) race and (perception of) qualifications are genuinely separable, given that "race" does not refer to mere phenotypic features by which race is identified, but rather to a broad social category that is not separable from other social categories. The problem with the causal analysis is that it makes it seem like the employer is discriminating insofar as they treat two candidates differently despite being identical ("**similarly situated**") in every respect except for race. But perhaps the social meanings of purportedly identical qualifications (e.g. a college diploma) are not independent of race. Kohler-Hausmann's point is that even if the qualifications of differently raced individuals are *not* the same (as a result of the same facially-neutral credentials having different meanings based on one's race), it is *still* discriminatory to treat them differently. But this relies on a normative argument about how people should be treated, not on a normatively-neutral causal claim about the effects of race.

I see Kohler-Hausmann (and collaborators) as posing the most developed argument that social categories such as race cannot be **disentangled** from other properties of individuals. One question to raise against this concern is how much it generalizes. The mere possibility that race and qualifications might be socially and psychologically inseparable does not show that they are, and certainly does not show that there are not factors that can be separated. Additionally, even if in principle one cannot treat social categories as entirely independent, perhaps in some contexts they are independent enough that one can legitimately make the **idealization** that they are. A further question is whether the concern ultimately has implications for how discrimination is to be tested. As I read her, Kohler-Hausmann is not in fact rejecting audit studies as a means of testing discrimination, but rather rejecting an incorrect interpretation of why it is that they are able to do so. So, at least in this case, it is not clear that the conceptual shift offered by her account makes an experimental difference. Audit studies are of course limited, and it is very plausible that the account could inspire new directions in testing for discrimination. But for now it is safe to say that this direction remains underdeveloped.

Connecting this to an earlier week, Mallon's discussion contains a basis for rejecting one of Kohler-Hausmann's arguments for why race's being socially constructed entails that in *cannot* be manipulated (contra Sen and Wasow). She argues that to test the effects of race by manipulating certain of its superficial manifestations, one need to presuppose that race just is those manifestations. So, for example, to test for the effects of race by manipulating skin color entails that race just is skin color, which is incompatible with social constructivism. If, however, one adopts the version of constructivism that Mallon labels the "folk-objectivist" account, one can reject the presupposition of the whole argument. The central/indicative distinction allows that one could test for the effects of race by manipulating indicators without claiming that it is nothing except for those indicators. In Weinberger (2022b, §5) I make a similar argument, but I had not at the time read Mallon.

### 4.2 Week 2: Causal Mediation Methods

In the causal model above, race influences being hired both directly as well as indirectly via qualifications. Facts about how a cause influences its effect are known as **path-specific effects**, where the influence via the direct path is known as the *direct effect* and that via the indirect path is the *indirect effect*. Such path-specific effects are studied using **causal mediation methods**, which increasingly play a role in studies of discrimination and fairness (e.g. Zhang and Bareinboim, 2018; Chiappa, 2019). In class, I started with a short presentation on Pearl's "Direct and Indirect Effects" (2001). In contexts with possible interaction between the cause and the intermediate variable (the *mediator*), the definitions of direct and indirect effects become fairly subtle. One must distinguish between "natural" and "controlled" versions of the direct effect and even when focusing on the natural direct effect (or the indirect effect), the direct effect of being black (vs. white) is different from that of the direct effect of being white (vs. black) (Weinberger, 2019). The upshot of all of this is that whether discrimination should be measured using path-specific effects (as well as *which* effects are relevant) calls for sustained philosophical analysis.

In her blog post "direct effects", Lily Hu argues against the interpretation of discrimination as a direct effect. Her piece provided an opportunity to revisit some of Kohler-Hausmann's arguments in the context of mediation methods. While there are opportunities for questioning whether she presents the most charitable reading of the "direct effect" view, the truth is that the scholars appealing to such a view themselves have yet to present a systematic presentation or defense of it. After discussing Hu, we went on to discuss my "signal-manipulation" proposal (Weinberger, 2022b). There I try to clarify what's at stake in debates over whether race is a cause, and argue that it is sometimes justifiable to include "race" as a causal variable in modeling a discrimination experiment even if the variable does not represent race in all of its sociological complexity. Just as in science in general, it is often legitimate to make the idealization that the activities of a local subsystem are largely independent of the rest of the universe, so too we should consider making the idealization that race can be disentangled from other social categories when testing discrimination. Whether such an idealization can be justified is a central question of the course as a whole.

## 5 Models and Idealization

A central topic within the philosophy of science is how models can play a legitimate role in science despite making idealized assumptions that we know to be false. In the context of discrimination, focusing on the topic of idealization gave us an opportunity to consider the way that discrimination is modeled within economics. We began with Sugden's well-known "credible worlds" paper (Sugden, 2000). His first example is the Schelling segregation model, which is a simple agent-based model of obvious relevance for discrimination. His second example – Akerloff's market for lemons paper – is also very relevant, since it concerns the notion of relying on **proxies** in information-poor environments. Proxies matter for discrimination both because it is illegal to use protected categories as proxies, and because it is unclear which proxies one is permitted to use – especially in algorithmic contexts.

We then turned to a preprint of Zaynep Pamuk's yet-to-be published "Rationalizing Discrimination", covering the history of Gary Becker's distinction between **taste-based** and **statistical** discrimination. Unlike taste-based discrimination, which is understood to be based on an irrational racial animus, statistical discrimination is rational in the sense that it involves using protected categories as reliable proxies to save one the time and effort of collecting more data. Pamuk goes through the history of the way that discrimination has been empirically modeled in econometrics, and notes how the assumption that the proxies used in statistical discrimination are *in fact* reliable is typically presupposed rather than defended. Moreover, in cases where one uses unreliable proxies reflecting problematic stereotypes, the appeal to such proxies might itself reflect personal prejudice. What makes the article particularly interesting isn't so much that Pamuk points to an unquestioned assumption in econometric models, but rather that her detailed discussion of the history of these models conveys how the discipline is structured in a way that determines which questions get asked and how they are studied. Since we didn't otherwise cover economics in this course, this article was very useful for giving students a sense of how economic approaches work, and was further helpful for disambiguating the senses in which discrimination might be labeled rational or irrational.

I included an article by Douglas Massey (1990) on the history of segregation in the US as a supple-

mental reading. As much as I love the intellectual questions stimulated by the Schelling segregation model, I didn't want the students to walk away from the class not being aware that segregation in the US is the result of centuries of intentional policies with continuing impacts.

# 6 Empirical Methods

## 6.1 Week 1: Causal Modeling

To introduce causal modeling, I assigned the chapter on causality from Barocas et al. (2023). I like how they cover both the graphical and potential outcomes frameworks and find that in general they are good at discussing the relevance of causal models for discrimination without overselling them. Nevertheless, it was a mistake to assign this chapter for introducing causal models, since it was just too big a learning curve for people without some technical background. In the end I just ended up doing a presentation of the key concepts I wanted them to know, including Simpson's paradox, d-separation, the causal Markov condition, and causal identifiability (confounding). We also discussed the paper by Bright et al. (2016) on causally interpreting intersectionality, which allowed us to discuss structural equations, revisit the notion of causal interaction, and to return to the theme of how to determine when a precise account of an imprecise concept should be considered adequate.

## 6.2 Week 2: Testing Police Discrimination

This week we covered the debate between Neil and Winship (2019) and Weinberger (2022a) on testing for police discrimination. Neil and Winship's discussion of benchmark statistics highlights how careful one must be when interpreting statistical data purporting to show discrimination or its absence. Such statistics are highly non-robust to the specification of additional information about the modeled scenario. This raises the question of what types of assumptions are needed in order to reliably infer discrimination from statistics. Neil and Winship believe that one can do so using the "similarly situated" criterion (see §4.1), which they seem to think is a purely statistical basis for differentiating the statistics that are and are not relevant to discrimination. In contrast, I argue that, given **Simpson's paradox**, there is no purely statistical sense in which certain benchmarks are privileged and one therefore requires some non-statistical assumptions to specify the target one is trying to measure when measuring discrimination. I then argue that causal models can do so. To be clear, I do not present causal models as *substitute* for normative theorizing, but rather claim that given the normative assumptions required for conceptualizing discrimination, causal models are useful for empirically testing discrimination claims. Crucially, while some people think that causal knowledge is only relevant when one plans to experimentally intervene, my argument points to a substantive role of causal assumptions in interpreting the data.

# 7 Discrimination in the Law

## 7.1 Week 1: Disparate Treatment and Disparate Impact

"Big data's disparate impact" (Barocas and Selbst, 2016) covers both the background for legal discussions of discrimination in the U.S. and the extension of discrimination law to algorithmic contexts. As such it served as an introduction to the remaining sections of the course. On the legal side, the most important distinction is that between **disparate treatment** and **disparate impact** (called direct/indirect discrimination in European contexts). Whereas disparate treatment refers to the direct use of a protected category in discrimination, disparate impact refers to the use of "facially-neutral" policy that does not explicitly take such categories into account, but whose implementation will lead to the unnecessary perpetuation of disparities. Over time, U.S. courts have become generally more skeptical of appeals to disparate impact doctrine. Additionally, discourse over affirmative action suggests that American society as a whole has become less tolerant of anti-discrimination policies that appeal to addressing past and ongoing injustices rather than to procedural fairness. Nevertheless, Barocas and Selbst compellingly argue that disparate treatment doctrine lacks the tools for dealing with algorithmic fairness, so it is unclear what legal tools remain given the demise of disparate impact.

One of the big takeaways of the article is that when it comes to algorithmic fairness, intentional discrimination is not the primary concern. Even if the designers of an algorithm do not seek to discriminate and explicitly forbid the algorithm for using certain prohibited variables, algorithms have the ability to recover information that is equivalent to knowing the values of the prohibited variables via a sufficient number of imperfect proxies. This point is also made compellingly by Prince and Schwarcz (2019), which I included as a supplemental reading. The problem raised in the article remains open, though the fact that algorithmic discrimination does not primarily involve intentions counts in favor of developing causal approaches. Although Lily Hu (§4.2) claims that causal models presuppose that the direct effect corresponds to racial animus (which is one of the ways that disparate treatment is explicated), there is nothing in fact in the model that requires the direct influence to go via the discriminator's bad intentions, and this is a virtue of such models. This does not, of course, prove that causal models are able to provide a solution, though this is an ongoing area of research (Plecko and Bareinboim, 2022).

## 7.2 Week 2: Zatz' Causal Legal Analysis of Discrimination

While it is increasingly common to propose causal mediation models for discrimination and fairness and to claim that direct and indirect discrimination correspond to direct and indirect effects, to my knowledge no one has actually given a thorough argument that the causal quantities in the relevant models in fact correspond to the intended legal notions of discrimination. Although it is not widely known, I view "Disparate impact and the unity of equality law" by Noah Zatz (2017) as the most promising basis for spelling out the legal interpretation under which disparate treatment and disparate impact correspond to direct and indirect effects, respectively. Although the discussion is not grounded in an up-to-date causal analysis, his discussion of causation generally gets things right. For instance, he makes clear that even when causal inference relies on population-level differences to establish causal effects, these average effects are still relevant to the individuals in the populations (even if we don't know precisely the individuals in which the effect occurs). The central argument in the paper relies on treating disparate impact and non-accommodation law as having a common basis. In non-accommodation law, an employer can be responsible for discriminating by making decisions that lead to barriers for individuals even without explicitly intending to harm those individuals. Similarly, with disparate impact employers are responsible for increasing a disparity that resulted from prior discrimination without intending to do so. In this way, Zatz motivates understanding disparate impact as corresponding to an indirect effect.

Zatz' article is important not just for those who defend the use of mediation models for analyzing discrimination, but also for those criticizing them. The weak point of his article comes in his discussion of *which* disparities one should treat as resulting from race in cases of disparate impact. He argues that because an individuals' race cannot be an effect of the variables of interest (e.g. education), confounding is not an issue, and we can treat the correlation between race and those variables as identifying an effect. This implies that all racial disparities count as discriminatory. This is worrisome, since one would have thought that the whole use of causal analyses of discrimination should be to provide a basis for differentiating discriminatory and non-discriminatory disparities. This issue therefore needs to be resolved before employing the account.

## 7.3 Week 3: Bostock v. Clayton County

If you're the type of philosopher who gets a thrill out of seeing philosophical debates have an impact outside of philosophy, you'll love the legal debate surrounding the Supreme Court decision *Bostock v. Clayton County* (2020). There the question was whether a man who was fired for being married to a man was discriminated against. On textualist grounds, the majority opinion argued that he was. The reasoning was that because had Bostock been a *woman* married to a man, she would not have been fired, the firing counts as discrimination on the basis of sex according to Title VII of the civil rights act. Accordingly, no further protection against sexual orientation discrimination must be independently legislated. As many, including Dembroff and Kohler-Hausmann (2022) have argued, the counterfactual appealed to by the majority was not the only plausible one. Had Bostock been a woman, but still been in a same sex relationship, she would still have been fired. So it seems like the counterfactual test employed is not suitable for resolving whether sexual orientation discrimination counts as sex discrimination, since the result one gets depends on which counterfactual one considers.

The Dembroff and Kohler-Hausmann article was useful for providing some background about the philosophical analysis of counterfactuals as well as for reinforcing Kohler-Hausmann's (2018) argument that questions about discrimination are not to be resolved by value-neutral counterfactuals, but rather call for substantive normative inquiry. The discussion also gave us an opportunity to clarify how Kohler-Hausmann would respond to my claim in Weinberger (2022b) that questions about what makes discrimination wrong can be dealt with independently of whether, given ones answers to those questions, a particular action is discriminatory. She would claim that just as in tort law, the question is never just whether damage was done, but rather whether damage was done negligently, so too with discrimination the question is never just whether one was treated differently, but rather whether one was treated differently in an illegitimate way. Accordingly, the descriptive and normative can never come apart.

Eidelson (2021) provides a clever textualist defense of the Bostock decision. His argument relies on the claim that the distinction between sex and a particular sex (e.g. being male) is metaphysically that between a determinable and its determinates and that it is possible to show that Bostock's sex understood as a determinable was a basis for his being fired without appeal to a counterfactual about how he would have been treated had he had a different determinate sex. Here I will not further go into the details of the argument, which provides much food for thought for those thinking about causation, counterfactuals, and discrimination.

# 8 Algorithmic Fairness

We began our discussion of algorithmic fairness with a response to Barocas and Selbst (2016) by Grimmelmann and Westreich (2016). This paper served as a useful pedagogical exercise, since although one can use it to develop a defensible argument, the conclusion drawn by the authors is overly general. The defensible argument is as follows. Barocas and Selbst claim that if an algorithm makes its decision by predicting a factor that is accepted to be legitimate for decision making, and that factor is correlated with race, then the only legal basis for treating that decision as discriminatory is disparate impact doctrine (which is increasingly disfavored). Grimmelmann and Westreich point out that there is something such an algorithm could be doing that would make it clearly discriminatory under disparate treatment. Namely, if the algorithm's predictive power results from using protected categories as *proxies* for the legitimate factor, this would be clearly illegal. Their point is that in order for the algorithm user to prove that what they are doing is not discriminatory, they must clarify what it is doing sufficiently to establish that it is not using protected categories as a proxy in this manner. The way that their conclusion overgeneralizes is that they make a more general claim that it must be possible to explain why the inputs used by the algorithm are relevant to the factor predicted. But as long as one can rule out proxies, it is unclear why any further demand for explanation is legally grounded.

As an exercise, I asked the class to use **d-separation** to develop a fairness criterion on the basis of Grimmelmann and Westreich's argument, and they were able to do so. The argument shows that an input that predicts a legitimate factor is discriminatory if the input and the factor are d-separated by a variable denoting a protected category. I don't see this criterion as being very useful in practice, since presumably there will be many causal paths between the input and the factor, and thus that the algorithm will be able to make its prediction without using a protected category as a proxy. The discussion was nevertheless useful for illustrating how one might begin to build an analysis of algorithmic fairness using causal notions such as d-separation.

Finally, we ended the class with a discussion of Creel and Hellman (2022). They argue that arbitrariness in decision making is not by itself problematic, but only becomes problematic at scale. So it is not necessarily a moral problem if a person doesn't get a job for an arbitrary reason, but if the same set of people are systematically not getting jobs across contexts, this is problematic. The worry about algorithms is that if the same ones get employed across many contexts, this increases the probability of systematic arbitrariness of this sort. In addition to being a plausible claim, it ties in to a discussion we were having since the first class about socially salient categories. A plausible story for why socially salient categories such as race and gender call for special protections is that the way these categories affect individuals is systematic across many contexts. Creel and Hellman's discussion helps clarify why systematicity would make a moral difference.

# References


Barocas, S., M. Hardt, and A. Narayanan (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Barocas, S. and A. D. Selbst (2016). Big data's disparate impact. *California law review*, 671–732.

Bright, L. K., D. Malinsky, and M. Thompson (2016). Causally interpreting intersectionality theory. *Philosophy of Science 83*(1), 60–81.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 7801–7808.

Creel, K. and D. Hellman (2022). The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy 52*(1), 26–43.

Dembroff, R. and I. Kohler-Hausmann (2022). Supreme confusion about causality at the supreme court. *CUNY L. Rev. 25*, 57.

Eidelson, B. (2021). Dimensional disparate treatment. *S. Cal. L. Rev. 95*, 785.

Grimmelmann, J. and D. Westreich (2016). Incomprehensible discrimination. *Calif. L. Rev. Circuit 7*, 164.

Haslanger, S. (2016). What is a (social) structural explanation? *Philosophical Studies 173*, 113–130.

Khalifa, K. and R. Lauer (2021). Do the social sciences vindicate race's reality? *Philosophers 21*(21).

Kohler-Hausmann, I. (2018). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev. 113*, 1163.

Mallon, R. (2004). Passing, traveling and reality: Social constructionism and the metaphysics of race. *Noûs 38*(4), 644–673.

Massey, D. S. (1990). American apartheid: Segregation and the making of the underclass. *American journal of sociology 96*(2), 329–357.

Neil, R. and C. Winship (2019). Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology 2*, 73–98.

Pearl, J. (2001). Direct and Indirect Effects. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420.

Plecko, D. and E. Bareinboim (2022). Causal fairness analysis. *arXiv preprint arXiv:2207.11385*.

Prince, A. E. and D. Schwarcz (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev. 105*, 1257.

Root, M. (2000). How we divide the world. *Philosophy of Science 67*(S3), S628–S639.

Ross, L. N. (2023). What is social structural explanation? a causal account. *Noûs*.

Sen, M. and O. Wasow (2016). Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science 19*, 499–522.

Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. *Journal of economic methodology 7*(1), 1–31.

Weinberger, N. (2019). Path-specific effects. *The British Journal for the Philosophy of Science 70*(1), 53–76.

Weinberger, N. (2022a). The insufficiency of statistics for detecting racial discrimination by police.

Weinberger, N. (2022b). Signal manipulation and the causal analysis of racial discrimination.

Zatz, N. D. (2017). Disparate impact and the unity of equality law. *BUL Rev. 97*, 1357.

Zhang, J. and E. Bareinboim (2018). Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.