Lily Hu[†]

What is "Race" in Algorithmic Discrimination on the Basis of Race?

Abstract: Machine learning algorithms bring out an under-appreciated puzzle of discrimination, namely, figuring when a decision made on the basis of a factor *correlated with race* is a decision made on the basis of *race*. I argue that prevailing approaches, which are based in identifying and then distinguishing among causal effects of race, in their metaphysical timidity, fail to get off the ground. I suggest, instead, that adopting a constructivist theory of race answers this puzzle in a principled manner. On what I call a "thick constructivist" account of race, to be raced is to be socially positioned in the way indicated by a certain set of statistical regularities on the basis of particular phenotypic traits. A thick constructivist sees that acting on the basis of correlations that constitute race qua social position *just is* acting on the basis of race, because races *just are* social positions that subject their member individuals to a particular matrix of social relations that define the raced position. This conclusion has considerable ramifications for our understanding of discrimination, algorithmic and beyond.

1. Introduction

In August 2018, the United States Department of Housing and Urban Development (HUD) filed a Housing Discrimination Complaint against Facebook, alleging that the company's algorithm-based advertisement system "mines extensive user data and classifies its users based on protected characteristics" and thereby "unlawfully discriminates by enabling advertisers to restrict which Facebook users receive housing-related ads based on race, color, religion, sex, familial status, national origin and disability."[1] In its response to the Complaint, Facebook pointed out that its machine learning system does not make use of a "Race" feature and as such does not even have the *ability* to target individuals based on race. But the line was unconvincing. In a formal charge the following year, HUD readily admitted the exclusion of protected class attributes from Facebook's ad system but claimed that it nevertheless discriminated on the basis of those attributes:

> Respondent [Facebook] combines the data it has about user attributes and behavior on its platforms with data it obtains about user behavior on other websites and in the non-digital world. Respondent then uses machine learning and other prediction techniques to classify and group users so as to project each user's likely response to a given ad. In doing so, Respondent inevitably recreates groupings defined by their protected class… [B]y grouping users who 'like' similar pages (unrelated to housing) and presuming a shared interest or disinterest in housing-related advertisements,

[†] PhD Candidate, Harvard University, Department of Philosophy, Department of Applied Mathematics

[1] Department of Housing and Urban Development, 'Housing Discrimination Complaint,' 2018, p. 2. https://www.hud.gov/sites/dfiles/PIH/documents/HUD_01-18-0323_Complaint.pdf

> Respondent's mechanisms function just like an advertiser who intentionally targets or excludes users based on their protected class."[2]

To clarify the grounds of the Charge, separate out two questions: First, why *does* the algorithm reconstitute protected class groups? And second, why does this reconstitution count as *discrimination*?

Granting the accuracy of HUD's reconstruction of the ad mechanism's functioning, answering the first question is easy. So long as there really is an empirically observed relationship between an individual's protected class attributes and their likelihood of clicking on a given advertisement, a machine learning system, which probes all sorts of statistical regularities in order to make high-accuracy predictions, is likely to learn it. Since the algorithm draws on data that tracks those patterns, how the system comes to reconstitute protected social groups is no mystery—basic observations from sociology and about how machine learning works explain it easily enough.

The second question is much trickier: why should this series of inferences constitute discrimination on the basis of, say, race? The exchange between Facebook and HUD drills down on a critical point of contention in debates about algorithmic discrimination: what does it take for a data-based system to act *on* or *on the basis of* race? HUD proffers a negative proposition: an algorithm can discriminate on the basis of race, even when it does not have access to a "Race" feature. Could such a bold claim be true? What theory of discrimination, and what theory of race, would make it so?

This paper focuses on the question of when algorithms discriminate on the basis of race. My aim, however, is not to hold algorithms up as some test case that can adjudicate ongoing philosophical and legal debate on theories of wrongful discrimination, nor is it to provide a new such theory. Rather, I focus on elaborating the problem of "algorithmic discrimination" in order to add to our thinking about what racial discrimination *descriptively* is and, hopefully, to do so in a way that gives insight into why we often have reason to be concerned with it.

Data-based predictive tools like the ones at the center of the *HUD v. Facebook* case clarify an underappreciated problem for discrimination theory: figuring when a decision made on the basis of a *feature correlated with race* is a decision made on the basis of *race*. This question is not only relevant to debates about algorithms. A version of it has dogged philosophers and legal scholars for decades

[2] Department of Housing and Urban Development, '*HUD v. Facebook*: Charge of Discrimination,' 2019, pp. 5-6. https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf

under the descriptor "proxy discrimination," in which an attribute that is a "proxy" for race is used in decision-making.[3] Predictive machine learning systems cast this old problem in new light.

The primary aim of this paper is to present an analysis of discrimination on the basis of race that offers a principled response to the practical question of when decisions made on the basis of *features that are correlated with race* are decisions made on the basis of *race*. This I call the *puzzle of algorithmic discrimination*. I pursue a solution to this puzzle by elaborating not a distinctive account of discrimination but a distinctive, social constructivist account of *race*. Social constructivist ideas, though commonplace in work on the philosophy of race and feminist metaphysics, have not significantly informed work on discrimination. This theoretical gap is laid especially bare by the puzzle of algorithmic discrimination, because social categories powerfully shape the data that drive the predictive prowess of machine learning algorithms. The rise of prediction-based decision polices shows that theorists of discrimination must pursue a deeper examination of the categories on the basis of which agents discriminate.

The following section (§2) summarizes the puzzle of algorithmic discrimination. Most extant analyses of algorithmic discrimination posit that understanding how race *causes* certain outcomes is crucial to solving this puzzle. This approach is correct in respects but cannot by itself make substantial progress toward a principled response to the puzzle of algorithmic discrimination. A conception of race as a cause is best elaborated within a social constructivist picture of race. In §3, I embark on a constructivist interpretation of robust correlations with race. What I call "thick constructivism" posits that certain correlations with "Black" uncovered by machine learning procedures do not merely track the incidence of certain outcomes among Black individuals; rather, they disclose social facts that *define* the category "Black," that reveal *what it is* to be Black or what being Black socially *consists in*. The thick constructivist answer to the puzzle of algorithmic discrimination has significant normative ramifications, which I discuss in §4: first, calling into question the commonly-held distinction between direct and indirect discrimination, and second,

[3] There are multiple conflicting legal accounts of what precisely makes an attribute a proxy for race in cases of proxy discrimination. Some adopt a wide interpretation: all attributes that correlate with race may be proxies for race in cases of disparate impact (e.g., Larry Alexander, 'What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies,' *University of Pennsylvania Law Review* 141, no. 2 (1992), pp. 149-219). Others require that the proxy attribute be chosen *because* of its correlation with race (e.g., Anya E. R. Prince and David Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data,' *Iowa Law Review* 105, (2020), p. 1257). The latter view draws a distinction between cases in which race is "directly predictive" and cases in which race is "indirectly predictive" of some target. I argue in the rest of this paper that the grounds for such a distinction must articulate a theory of what race is as a social category, which most scholars writing on discrimination have not done.

unsettling the presumptive boundaries that circumscribe the wrongs of racial discrimination from the broader set of racial wrongs.[4]

2. Algorithmic Discrimination on the Basis of Race

The *HUD v. Facebook* case well illustrates a central dispute in cases of potential algorithmic discrimination, with each party taking a distinct interpretation of the notoriously vague "because of" race clause commonplace in accounts of discrimination. Consider the following, I think uncontroversial, non-moralized analysis of discrimination on the basis of race.[6]

> (RD) X *discriminates on the basis of race* if X treats Y differently from how X treats or would treat some Z because Y (supposedly) has racial status R and Z (supposedly) has a different racial status R´.

In the literature, "because" tends to be read in two ways.[7] On one reading, "because" refers to X's taking race as a *reason* for her differential treatment. This *reasons* interpretation of "because" asks us to look into the considerations X took in acting the way she did. For example, if a group of Facebook employees sought to target ads to individuals based on race and so designed the ad system to create racial groupings, the company would be discriminating on the basis of race on the reasons interpretation of (RD). On another reading, "because" refers to X's treatment of Y as being *causally* affected by race in a non-reasons way. This *causes* interpretation asks us to consider how race might have played a causal role in how X treated Y *aside from* being a part of X's reasons, perhaps even in a way unbeknownst to her. In making decisions via features that exhibit non-accidental correlations with race, machine learning tools would appear to meet the causal connection criterion, thereby discriminating on the basis of race according to the causes interpretation.[8]

[4] A similar line of argument, I believe, also applies for discrimination on the basis of sex and gender, though I will not pursue it here.

[6] This analysis is similar to the "Discriminator's Reasons Analysis" in John Gardner, 'III—Discrimination: The Good, the Bad, and the Wrongful,' *Proceedings of the Aristotelian Society* 118, no. 1 (2018), pp. 55-81. However, my analysis is explicitly constructed to include more than just a reasons-based analysis of discrimination. For another non-moralized analysis of discrimination, see Kasper Lippert-Rasmussen, *Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination* (Oxford: Oxford University Press, 2014).

[7] Patrick Shin discusses a causes versus reasons analysis relevant to U.S. discrimination law in 'Liability for Unconscious Discrimination: A Thought Experiment in the Theory of Employment Discrimination Law,' *Hastings Law Journal* 62, (2010), p. 67.

[8] Readers familiar with discrimination theory may wonder why I do not stick with the more common direct vs. indirect—or in U.S. jurisprudence, disparate treatment vs. disparate impact—distinction. I opt for the reasons vs. causes framework for three reasons. First, a conception of race as acting causally in our world is central to both technical work in algorithmic discrimination and constructivist accounts of race. Adopting this causal picture helps connect work in race constructivism with accounts of algorithmic discrimination on the basis of race. Second, I find the standard gloss on indirect discrimination to be somewhat elusive. For example, Sophia Moreau locates the distinction between direct and indirect discrimination "in the 'closeness' of the disadvantageous effects to what the agent has done." Something like a

To get to the puzzle of algorithmic discrimination, first set aside some easy cases. I gather that everyone would agree that acting on covert intentions to create racial groupings constitutes racial discrimination (and wrongful racial discrimination at that). So, for the sake of argument, I will grant that the ad delivery system is not intentionally designed to deliver less favorable advertisements to certain racial groups. An algorithm might also discriminate on the basis of race on the reasons interpretation by simply including the "Race" feature in its learning process. I tread carefully here because it is not so clear to me what it means for algorithms to take something as a reason. I will return to this key point in the following sections, but for now, I make the simplifying assumption that the inclusion of a "Race" feature in a machine learning process that generates decisions is the *only* way an algorithm acts "on the basis of" race on the reasons picture.[9] The tricky cases, then, are ones in which the algorithm is not designed with racial motivations and also does not include a "Race" feature—cases like *HUD v. Facebook*.

For these cases, the reasons versus causes framing exposes a deep fault line. Adopt the reasons interpretation and machine learning algorithms *never* discriminate on the basis of race; go with the causes interpretation and machine learning algorithms, if working properly, *necessarily* discriminate on the basis of race. Machine learning *just is* a process of uncovering associations among various events, outcomes, categories, and attributes. Highly accurate algorithmic predictions that non-accidentally correlate with race do so because the process successfully "learns" the social effects that racial distinctions have in the world and leverages these correlations in making predictions. It seems, then, that neither of these interpretations are—or at least neither of these *crude* interpretations—of reasons and causes are immediately helpful in passing judgment on presumptive cases of racial discrimination when the "Race" feature is excluded from an algorithmic system. If algorithms can indeed act on the basis of race *without* acting on the feature "Race," we will need to look beyond this divide for a descriptive analysis of discrimination on the basis of race that spells

causal connection to the "protected trait" seems the most natural way to interpret this "closeness." Finally, there are various ways race could causally contribute to disadvantaging treatment—via unconscious or implicit bias, epistemically justified inferences from racial statuses, or facially neutral policies that "happen to" align with racial inequalities. Speaking of race as operating causally in different ways allows us to elaborate these distinctions further if needed. Sophia Reibetanz Moreau, 'Equality and Discrimination,' in ed. John Tasioulas, *The Cambridge Companion to Philosophy of Law* (Cambridge: Cambridge University Press, 2020).

[9] I make this simplification to skirt around the need for a general account of acting "on the basis of" X (on the reasons view). However, later in the paper, my considered view does away with this assumption. I also set aside cases in which the "Race" feature is included in the algorithmic pipeline so to ensure that the algorithm *does not* "discriminate" on the basis of race. While I have no doubt that practitioners use features in this way, such cases assume that one *can* discriminate on the basis of race when the feature is not included, which is to presuppose a solution to the puzzle of algorithmic discrimination. In trying to *come to* a solution, I think it fine to set aside these cases.

out, in a principled manner, how a decision made on the basis of features correlated with race *could* be a decision made on the basis of race.

Scholarship on algorithmic discrimination has overwhelmingly advanced a version of the causes interpretation.[10] The causes interpretation of racial discrimination shifts from foregrounding the primarily moral concern that one's race might elicit bad racially-motivated responses, instead foregrounding the primarily political concern that race and racial categories might have broad consequences on the fairness of our basic institutions. This approach is thus attractive for those who see the problem of racial discrimination as systemic in nature; as such, it is particularly well-suited to the new frontier of algorithmic decision tools, which, by predicting future outcomes using data from the past, can either reinscribe or ameliorate existing patterns of inequality.

But pursuing this route toward resolving the puzzle of algorithmic discrimination comes with challenges that cannot be resolved by appealing only to resources available within the causal framework. For one, an expanded analysis of racial discrimination that counts *all* decisions made on the basis of factors causally linked to race as decisions made on the basis of race is implausibly permissive. Ruling that decision-making on the basis of *any* such factors constitutes discrimination is an outcome that few, even those sympathetic to the causes interpretation, would likely accept. So, we will need a couple of things from a successful causes interpretation answer to the puzzle of algorithmic discrimination. Faced with a complex web of causes, we need an account of which causal relations with race do and do not count. And this account should be well-motivated so to explain why acting on a feature with *that kind* of causal connection to race should count as acting on the basis of race.

Some technical work in causal approaches to discrimination has already ventured to provide responses to these questions. Though a full discussion of these proposals is beyond the scope of this

[10] One reason for convergence on the causes interpretation is that it is more amenable to those mathematical approaches central in technical work on algorithmic discrimination. In formalizing the causes analysis, computer scientists have developed their own theories of *proxies*. Some works explicitly note that proxy attributes need not have a causal connection but may be "just correlated" with a given protected attribute. This distinction, however, strikes me as misleading, since it is usually not meant to suggest that there is no causal story *at all* to tell linking the proxy and protected attribute. To avoid this confusion, I count all attributes that exhibit a non-accidental correlation with race as being causally connected to race. Relatedly, some work in algorithmic fairness considers all data systems that make use of proxies to be engaging in proxy discrimination, implying that any causal influence of race on a decision via proxies is prima facie worrisome. My non-moralized descriptive analysis does not weigh in on this judgment. For an article in the algorithmic fairness literature that defines proxies and proxy discrimination in this way, see Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf, 'Avoiding Discrimination Through Causal Reasoning,' *Advances in Neural Information Processing Systems*, (2017), pp. 656-666.

paper, a few remarks will hopefully highlight where I believe them to fall short.[11] One prominent strand of research suggests that we first draw a diagram showing how race and other features in the system are causally related to each other and then designate which causal pathways are and are not discriminatory.[12] Causal pathways deemed discriminatory must have their effects struck out and cannot contribute to the algorithm's computations. The problem with this approach is the *grounds* one needs for identifying suspect pathways cannot solely come from within the causal framework. Most work in the area concedes as much and submits that the task of designating discriminatory pathways must draw on normative considerations outright. But this suggestion is simply unresponsive to the puzzle of algorithmic discrimination, which calls for a *descriptive* account of what it is to act on race, setting aside any *particular moral* theory about why acting in that way might be wrongful. In skipping to a moralized conception of discrimination, these accounts overlook the question of what it is to act on the basis of race period, thus ignoring precisely what makes algorithms so compelling as potential agents of discrimination and failing to approach the central disagreement in cases of disputed algorithmic discrimination: did the algorithm act on the basis of race even though it lacked a "Race" feature?[13] Furthermore, any such method that *does* aim to provide a descriptive account is only able to pick out those pathways that count as acting on the basis of race in light of a fundamental theory of race. But causation-based approaches are largely silent on these matters of metaphysics. In fact, if we can read any theory of race into these approaches, it is one that sees strict *distinctions* between race and those features that are non-accidentally correlated with it. Acting on features with certain causal pathways to race would seem to be precisely a case of *not* acting on race.

Causal counterfactual approaches to algorithmic discrimination suffer from similar silences: in virtue of what facts do we decide what the relevant counterfactuals are or what the counterparts

[11] From a completely different line of critique, Issa Kohler-Hausmann and I have argued that so long as these accounts hew closely to the methods of causal inference in discussing how race acts causally, they face significant metaphysical and epistemological challenges; see 'What's Sex Got To Do With Fair Machine Learning?,' *Proceedings of the 3rd ACM Conference on Fairness Accountability and Transparency*, (2020).

[12] See e.g., Razieh Nabi and Ilya Shpitser, 'Fair Inference on Outcomes,' *Proceedings of the AAAI Conference on Artificial Intelligence* 32, (2018), pp. 1931-1940; Silvia Chiappa and Thomas P. S. Gillam, 'Path-Specific Counterfactual Fairness,' *Proceedings of the AAAI Conference on Artificial Intelligence* 33, (2019), pp. 7801-7808; Razieh Nabi, Daniel Malinsky, and Ilya Shpitser, 'Learning Optimal Fair Policies,' *Proceedings of Machine Learning Research* 97, (2019), p. 4674.

[13] A skeptic of this formulation might argue that discrimination is instead entirely a matter of certain unequal *outcomes* between groups. This is simply a different account of the concept of discrimination, one that is orthogonal to what I take to be the central disagreement in cases of disputed algorithmic discrimination: disagreement about when a machine learning system acts on race in the first place.

of certain features in other (non-discriminatory) possible worlds are?[14] There, too, a theory of race is required to even get the method off the ground. Noting that race stands in an important causal relationship with a given feature does not by itself make any headway in determining whether acting on that feature is acting on race—unless, that is, the observation *works from* an account of race, *in light of which* that feature and causal relation is privileged over others.

The principal issue with causation-based approaches, then, is not *just* that facts about causation alone do not give us enough information to pick out those non-"Race" features that count from those that do not count. Instead, it is that any implications such theories have for solving the puzzle of algorithmic discrimination will depend entirely on the theory of race that they posit—a theory that the approaches themselves cannot deliver. Disagreement among candidate solutions stems from disagreements about what race as a social category *is*. Taking metaphysics seriously is by no means self-indulgent.

Yet causal approaches are suggestive of an account of race. In highlighting a picture of race as producing effects in the world, they suggest that what race *does* is a key feature of what race *is*. I wholeheartedly endorse this proposal. But the fully elaborated account must focus on exactly what kind of a thing could be causally efficacious in the distinctive ways that race and racial categories are. I turn to these social metaphysical questions in the next section.

3. Constructivism about Race

To make sense of a causes interpretation of discrimination on the basis of race, we need to elaborate an analysis of race as a social category that is causally efficacious—as something that, in Charles Mills' words, "moves people."[15] But the straightforward assertion that "race acts as a cause" can suggest the rather unhelpful notion that race is just an individual attribute that triggers certain responses and not others. The thought goes something like this: the causal story of how Jamal's race caused the police officer to search him is given by a vignette of the police officer "perceiving" Jamal's race at time t_1 and responding to this racial status by taking a course of action that ends in his searching Jamal at some later time t_2. This series of events presents a causal story of sorts that links

[14] See e.g., Matt Kusner, Joshua Loftus, Chris Russell, Ricardo Silva, 'Counterfactual Fairness,' *Advances in Neural Information Processing Systems* 31, (2017), pp. 4066-4076; Yongkai Wu, Lu Zhang, and Xintao Wu, 'Counterfactual Fairness: Unidentification, Bound and Algorithm,' *Proceedings of the International Joint Conference on Artificial Intelligence* 28, (2019), pp. 10-16.

[15] Charles Mills, *Blackness Visible: Essays on Philosophy and Race* (Ithaca: Cornell University Press, 1998), p. xiv.

the racial status "Black" to a police search. But it leaves completely opaque the background conditions of the interaction: why "Black" had the causal effect that it had, with the police officer reacting the way he did, with *search* rather than *pass*. If a causes interpretation of discrimination is primarily concerned with the systemic nature of racial disadvantage, answers to these causal questions should explain why "Black" has causal capacities at all and why these capacities are so configured. If racial discrimination concerns a certain pattern of social responses to "race," we need a picture of what race is, not as an individual attribute but as a *social category* that structures social responses. An explanation for why Jamal's *being Black* "caused" the police officer to search Jamal must address the fact that the racial category Black is meaningful to the police officer in a way that structures his course of action. Perhaps in this case, the police officer responds the way he does because, for him, *Black* is linked to *danger* and *violence*. If so, we cannot understand how Black "causes" police search without understanding the origins of this representational content.

Social constructivism highlights that the myriad causal powers of Jamal's being raced Black are connected with the social kind of thing that being raced Black is and how the category is socially constructed as such. It behooves me to be explicit about what focusing on a constructivist account of race means for the account of racial discrimination I look to forward. Social constructivism does not posit a metaphysics of race with which other theories substantially disagree. As Ron Mallon notes, the three leading theories of race—racial skepticism, racial constructivism, and racial population naturalism—agree that racial concepts "causally affect persons in both superficial and profound ways."[16] That is, each accepts that most correlations observed (and algorithmically predicted) between race and socially important outcomes are due to some causal role that race plays in certain social circumstances.

This might lead us to wonder what a social constructivist account *does* get us. If not a distinctive metaphysical picture, a distinctive political one. Constructivists foreground the idea that racial categories are socially constructed to stand in the causal relations they do. Social and political forces continually shape the boundaries of racial categories, their meanings, and as a result, their causal roles.[17] Since "races" do not share essential characteristics that naturally mark them off from each other, racial divisions can only be maintained and produced anew by *race-making institutions*. For example, consider how the U.S. police and prison system function as institutions that produce and maintain race and racial difference. The expansion of the American carceral system beginning in the

[16] Ron Mallon, '"Race": Normative, Not Metaphysical or Semantic,' *Ethics* 116, no. 3 (2006), p. 545.
[17] Michael Omi and Howard Winant, *Racial Formation in the United States* (New York: Routledge, [1986] 2014).

late 1960s reinforced representations of Blacks as deviants, criminals, and dangerous.[18] The rise in "Black crime," itself a product of *race-laden* policies, further "justifies" policing practices such as racial profiling, such that Blackness becomes equated with criminality at both a micro-level of everyday social meanings and interactions and a macro-level of legitimized racial policing and policy-making rationale.[19] These effects spill over into other spheres such that Blacks are further socially, economically, and politically marginalized in highly specific ways. The concept of race is thereby further embedded into our society, figures into new and different non-accidental generalizations and asserts new and different pressures on Black lives. On the constructivist account of race, racial categories are what they are—and do what they do—because of the social and political processes that construct them as such.

3.1 "Thick" Constructivism about Race

One further way to flesh out the critical constructivist's project is to think of race as *hierarchically* constructed. Sally Haslanger takes race and gender to be social positions wherein certain characteristics of phenotype and ancestry (in the case of race) and physical sex traits (in the case of gender) "justify" differential treatment and positioning.[20] Seeing the race structure as hierarchical means seeing continuity in how Blacks are subordinated, even if the particulars of this subordination may vary by historical moment. The category "Black" is constructed in such a way that Black individuals occupy a social position defined by a set of norms and expectations, of privilege and of injury, that is as a whole oppressive. Haslanger's approach directly connects racial oppression to many observed correlations with race: "[G]roup domination [and] the effects of earlier injustice position subordinate groups socially and economically so that their members have much more in common than their group membership."[21] Robust non-accidental correlations between outcomes of disadvantage and the "Black" label are simply manifestations of an unjust race structure at work.

Now, if races are constructed *as* different, then consistent non-accidental correlations with race are not just unfortunate statistical regularities that algorithms must find a way to skirt around in

[18] Loïc Wacquant, 'Deadly Symbiosis: When Ghetto and Prison Meet and Mesh,' *Punishment & Society* 3, no. 1 (2001), pp. 95-133; 'From Slavery to Mass Incarceration: Rethinking the "Race Question" in the US,' in Donaldo Macedo and Panayota Gounari (eds.), *Globalization of Racism* (New York: Routledge, 2016), pp. 94-110.

[19] Robert Lieberman uses the term *race-laden* to describe policies that perpetrate racial injustice that are instituted without any intention that they do so in *Shifting the Color Line: Race and the American Welfare State* (Cambridge: Harvard University Press, 1998).

[20] Sally Haslanger, 'Gender and Race: (What) Are They? (What) Do We Want Them to Be?', *Noûs* 34, no. 1 (2000), pp. 31-55.

[21] Haslanger, 'Oppressions: Racial and Other,' p. 324.

order not to discriminate. They track empirical facts descriptive of what races qua social positions *are* and what it is to be raced *in the first instance*. I call accounts that adopt this view "*thick* constructivist" accounts of race, in contrast with "*thin* constructivist" accounts, which focus on races as persons sharing sets of phenotypic features, supposedly linked to geographic ancestry, that trigger differential social treatment.[22] For thick constructivists, the racial category Black is defined in reference to its overall social position of disadvantage: to be Black is to be marked, on the basis of one's phenotype and supposed ancestry, for certain norms, expectations, and practices of subordination.[23] Different accounts may be further distinguished by the subordinating social factors emphasized in defining the category. For example, consider Du Bois' response in *Dusk of Dawn* to his white interlocutor who asks, "[W]hat is this group; and how do you differentiate it; and how can you call it 'black' when you admit it is not black?" Du Bois says, "[T]he Black man is a person who must ride 'Jim Crow' in Georgia," defining Black by its relationship to legal institutions of racial subordination.[24] Adrian Piper offers an account wherein the *social experience* of "being Black" defines the racial category, an experience that is ultimately dictated by the fact of a "white racist society" serving "punitive and damaging effects."[25] Expanding a thick constructivist-like approach to the race concept more generally, Paul Taylor defines races as "probabilistically defined populations that result from the white supremacist determination to link appearance and ancestry with social location and life chances."[26]

Taylor's theory, with its mathematics-inflected language, speaks most directly to the "correlations with race" issue at the heart of the algorithmic discrimination puzzle. If to be raced Black is to stand in a particular subordinated relation to social forces of white supremacy that dramatically influence one's life chances, then the social grouping "Black" will be, as Taylor puts it, a "probabilistically defined population" characterized by a certain distribution over life chances resulting from the application of those forces on individuals within the group. Note that race as a

[22] My "thick" and "thin" distinction is inspired by Mallon who describes "thin" constructivism as taking race to only refer to classifications of morphology and ancestry that have, as it so happens, become socially significant in '"Race": Normative, Not Metaphysical or Semantic,' p. 535.

[23] My constructivist account of race constructivism should not be confused with what Kwame Anthony Appiah distinguishes as *racial identification*. Identification is a matter of an individual's or group's self-conceptualization by "reference to available labels, available identities." I want to maintain a distinction between accounts of race and accounts of racial identity. My primary interest in this paper is the first. Kwame Anthony Appiah, 'Race, Culture, Identity: Misunderstood Connections,' in K. Anthony Appiah and Amy Gutmann, *Color Conscious: The Political Morality of Race* (Princeton: Princeton University Press, 1998), p. 78.

[24] W. E. B. Du Bois, *Dusk of Dawn: An Essay Toward an Autobiography of a Race Concept* (New York, 1940; reprint, New Brunswick, N.J., 1992), p. 153.

[25] Adrian Piper, 'Passing for White, Passing for Black,' *Transition* 58, (1992), p. 30.

[26] Paul C. Taylor, *Race: A Philosophical Introduction,* 2nd edn (Cambridge, UK: Polity, 2013), pp. 89-90.

causal factor in individuals' lives definitely plays a role here—being marked "Black" (causally) subjects Black individuals to certain types of social treatment—but this constructivist account of the category is a constitutive one: "Black" is defined in terms of the social position that its members occupy such that to be "Black" is to stand "in some relation to social agents or social factors."[27] For example, one might partly define *being Black* in terms of one's relation to institutions of policing by virtue of facts about one's phenotype and ancestry, which has an overall effect of subordination. Taking this view, we would *expect* policing data to be racially skewed. A positive correlation between the features "Black" and "Past Searches By Police" does not expose just a causal effect of one's being marked with the attribute "Black," but rather a descriptive fact about what Black *is* as a social assignment. Racial profiling and racialized police brutality do not only reflect and reinforce pre-existing racial injustices. They forge a distinct pillar in the hierarchal race structure and partly form the thick social position that is the category Black.[28] More generally, thick constructivism denies a clean distinction between something understood as race R "itself" and the complex of social factors that constitute the thick position of race R.

We are now coming close to drawing thick constructivist conclusions about algorithmic discrimination, but I want to first address two points about this interpretation of correlations with race that warrant further treatment before discussing the normative upshot of the approach. First, a thick constructivist account of race need not take *all* non-accidental correlations with the feature "Black" to be constitutive elements of the racial category Black. Some—in fact I would posit, most—causal effects of being labeled "Black" and encountering the world as Black are not defining features of the category. The question of what is essential to being Black is an important conceptual, and I would suggest political, question that is not reducible to data-mining. The question of what counts as partly constitutive of race R does not map cleanly onto, say, the robustness or strength of a feature's correlation with "Race R."

This might be surprising, as it suggests that weaker correlations with "Race R" or correlations that exist in a narrower range of cases might indicate social facts "more essential" to

[27] Esa Díaz-León, 'What is Social Construction?,' *European Journal of Philosophy* 23, no. 4 (2015), p. 1142. Sally Haslanger contrasts constitutive social construction from causal social construction in 'Social Construction: The "Debunking" Project,' in S. Haslanger, *Resisting Reality: Social Construction and Social Critique* (Oxford: Oxford University Press, 2012), pp. 113-137.

[28] My suggestion that racialized policing practices are a pillar of the hierarchical race structure is agreeable to Annabelle Lever's observation that what makes profiling against Blacks, rather than whites, distinctively morally problematic has something to do with what distinguishes the racial category Black from the category white. 'Racial Profiling and the Political Philosophy of Race,' in ed. Naomi Zack, *The Oxford Handbook of Philosophy and Race* (Oxford: Oxford University Press, 2017).

what it is to be raced R than stronger and more broadly true correlations. But this is the correct conclusion. Thick social constructivist accounts emphasize the practical political utility of our concepts. A good analysis of race is one that highlights those urgent features of our unjust raced society—hence, the tendency for thick constructivists to include things like relations of privilege tied to whiteness and relations of subordination tied to Blackness. Further, constructivists look to provide analyses of our categories that shed light on a host of social phenomena. Accordingly, concepts and categories should be judged in part by their explanatory power. Since the strength and robustness of a correlation does not necessarily reveal anything about its practical or theoretical usefulness, there is no reason to expect these measures alone to determine whether some factor is essential to the social category.

This brings me to a second potential worry for a thick constructivist reading of correlations with race. If some social factor is essential to *being Black*, doesn't the feature need to be perfectly correlated with the racial label "Black"? For any proposed defining feature, though, an objector might show that it does not exhibit a perfect correlation with the label "Black" and therefore cannot be an essential feature after all, since someone who is Black does not have it. This objection is a version of the *commonality problem*, which questions whether there is any feature true of all members of a social category in all societies, past and present.[29] Debate about the commonality problem is far from settled, but I do not think it precludes applying a thick constructivist lens to correlations with race. First, part of the worry can be mitigated by ensuring that thick constructivist analyses of Black are not overly specific, pinpointing social factors of being Black that do not apply widely to Blacks in some way or another. But more importantly, an account of "Black" that refers to the hierarchical race structure need not require that all Black individuals experience particular *outcomes*. Even if *being Black* is defined as standing in a certain relation to policing that is overall subordinating, this does not mean that every Black individual experiences that relation similarly. Not all Blacks will be searched by police, nor will all Blacks personally experience the violence of a punitive carceral state. Many Blacks recorded in data sheets will have '0's or 'N/A's in columns counting previous encounters with police, but this does not mean that they do not suffer race-related harms from policing. Race-based harms and injustices of policing encompass more than individuals' direct encounters with police. For one, if race-based risks count as harming, current institutions of policing

[29] The commonality problem is a matter of ongoing debate in feminist theory. See, e.g., Alison Stone, 'Essentialism and Anti-Essentialism in Feminist Philosophy,' *Journal of Moral Philosophy* 1, no. 2 (2004), pp. 135-153; Theodore Bach, 'Gender is a Natural Kind with a Historical Essence,' *Ethics* 122, no. 2 (2012), pp. 231-272.

harm all Blacks probabilistically. This seems plausible to me, but even if you are wary of probabilistic harms, current policing practices surely influence how most Blacks navigate their lives—how they act in public, how they parent, whether they call police in emergencies—out of concern about what *would* happen if they were to encounter police. These are also race-based harms of policing. Finally, institutions of policing contribute to the stigmatization of Blacks as criminal, violent, and dangerous. These status harms and their material and psychic effects are borne by all Blacks.[30] Thus, the fact that correlations between "Black" and certain policing outcomes are not perfect neither shows that standing in a certain subordinating relation to policing cannot be a constitutive feature of being Black nor does it show that not all Blacks suffer injustice *as Blacks* at the hand of policing institutions.[31]

3.2. The Thick Constructivist Solution to the Puzzle of Algorithmic Discrimination

With these theoretical resources in place, I claim we *can* elaborate a causal analysis of discrimination on the basis of race that answers the puzzle of algorithmic discrimination—but we need to understand how race acts as a cause in the right way, an understanding advanced by a thick constructivist account of race. I argue that when it comes to algorithmic decision systems, discriminating on the basis of features correlated with race constitutes discrimination on the basis of race when those features are not merely causally related to but *constitutive of* the racial category qua thick social position.

Recall that thick constructivists define *being Black* in terms of a certain set of subordinating social factors. In the statistical language familiar to discussions of algorithmic discrimination, the category Black is defined as a social position whose occupants are probabilistically more likely to, as examples, be on the receiving end of a stop-and-frisk action or be a victim of state violence—on the basis of having a certain set of phenotypic traits, including skin color, hair texture, and so on. On a thick constructivist interpretation of algorithmic discrimination on the basis of race, an algorithm that includes the "Race" feature in its machine learning process and one that does not but does

[30] Appiah discusses probabilistic and identitarian status harms suffered by all Blacks in '"Group Rights" and Racial Affirmative Action,' *The Journal of Ethics* 15, no. 3 (2011), pp. 265-280.

[31] The observation that structures of anti-Black oppression do not harm only those Black individuals who directly interact with them has ramifications for how we evaluate quantified racial differences in certain social outcomes. For example, it is common social scientific practice to see the difference between, say, the frequency of police searches affecting white communities and the frequency of police searches affecting Black communities as approximating the police search harms suffered by Blacks qua Blacks. But if police search practices are a force of anti-Black oppression, they in some way or another impact *all* Blacks, not just those searched by police, a fact not captured by this numerical difference in frequency.

include features that exhibit non-accidental correlations tracking social facts constitutive of being raced R may *both* be discriminating on the basis of race R. Note that this rendering of the causes interpretation of discrimination on the basis of race is not automatically implausibly broad. On my view, drawing on attributes that are causally related to but not constitutive of race qua thick social position does not constitute discrimination on the basis of race.

The constructivist framework offers insight into the lament oft-repeated in discussions of algorithmic discrimination that an "algorithm can learn race even if you exclude race" (recall HUD's charge that Facebook's algorithm "inevitably recreates groupings defined by [race]").[32] The thought here seems to be that insofar as race leaves its mark on "non-race" features, keeping *those* features around in the machine learning pipeline will contribute to racially biased outcomes. This reasoning explains how algorithmic predictions can come to exhibit racial skew but does not give an account of *why* acting on the basis of those features should be considered acting on the basis of race. The thick constructivist interprets the claim that the algorithm "learns race" *quite literally*: it learns, via these correlations with race, what race *is*. To understand the feature "Race" that appears as a dataset column heading as marking the "real" social categorization of discriminatory concern is to mistakenly adhere to a data reification of the race concept. Acting on the basis of correlations produced by social facts that constitute race qua social position just is acting on the basis of race, because races just are social positions that subject their member individuals to the matrix of privileging and subordinating social relations that define what it is to be raced.

This might now look like a sleight-of-hand that moves the thick constructivist solution away from the causes interpretation toward the reasons interpretation of discrimination. Recall from §2 our simplifying assumption that a machine makes a decision "*because* of race" on the reasons view only when it includes the "Race" feature in its optimization process. The worry about adopting this narrow reasons interpretation, though, was that counting these as the only cases of discrimination on the basis of race would make charges of discrimination too easy to dodge—after all, it is easy to simply remove the "Race" feature from a dataset often to no effect on the algorithm's issued outcomes. But if, as the thick constructivist claims, features and correlations tracing social facts that constitute race qua thick social position *just are* what race is, then including and drawing on these features and correlations *is the same as* including and drawing on race. So, even on this highly

[32] Department of Housing and Urban Development, '*HUD v. Facebook*: Charge of Discrimination,' p. 5. For an example the claim that algorithms discriminate on the basis of traits that are excluded from the learning process, see, Betsy Anne Williams, Catherine F. Brooks, and Yotam Shmargad, 'How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications,' *Journal of Information Policy* 8, (2018), pp. 78-115.

simplified understanding of what it takes for a machine to take race as a reason, the algorithm discriminates on the basis of race. This suggests that excluding the "Race" feature is neither necessary nor sufficient to ensure that the system does not discriminate on the basis of race.[33] These observations look toward the need to elaborate other conceptions of acting on the basis of race on the reasons interpretation and to consider how machines can meet those standards. Constructivism, in tinkering with the "race" part of an analysis of discrimination on the basis of race, may show the need to reconfigure the "discrimination" part as well.

With the main parts of the account now on the table, a few clarifications and caveats are in order. The thick constructivist solution does not cover *all* the ways an algorithm can discriminate on the basis of race when it lacks a "Race" feature. I leave open the possibility that an algorithm could discriminate on the basis of race by drawing on features that, though *not* constitutive of any racial category, prove predictively useful *simply because* they correlate with the category. This situation commonly arises when race correlates with a predictive algorithm's target outcome, and the algorithm subsequently learns that drawing on features correlated with race is a successful predictive strategy. It seems to me that such cases can be covered by an expanded account of discrimination on the reasons interpretation, broader than the one on offer here, though a full defense of this position requires an articulation of how machines may take race as a reason.[34]

Throughout this paper, alongside my arguments about constructivism and race, I have tried to show how the puzzle of algorithmic discrimination reveals the need to take metaphysics seriously. The thick constructivist solution proposes that we must consider first those social relations and factors that form the distinctive position that is the racial category Black before we can come to an analysis of discrimination on the basis of being Black. Let me now clarify how we know which social facts these are and more generally, what theorizing about race consists in. On my view, the social metaphysics needed for this project is not based on *a priori* theorizing about concepts. Rather, a helpful thick constructivist analysis draws heavily on historical and empirical study of the various social relations of privilege and subordination that construct and uphold our actually existing

[33] Why not necessary? A full argument will require a theory of what it takes for an algorithm to take something as a reason—as well as an adequate understanding of how machine learning works (which we currently lack)—but it seems to me that the mere inclusion of a "Race" feature in a learning procedure does not guarantee that the algorithm will "make use" of it in a way that should count as an instance of its taking race as a reason.

[34] I thank David Gray Grant for pushing me to address cases of pure proxy discrimination.

hierarchical racial order. Only by scrutinizing these real-world social phenomena can we arrive at an accurate picture of what social facts are constitutive of the racial category Black.[35]

Adopting an empirically-informed metaphysics assuages a worry that insofar as the race structure is not our only hierarchical social structure, it will be difficult to distinguishes cases of discrimination on the basis of race from, say, cases of discrimination on the basis of sex. This is indeed a subtle issue. But it is one that social scientists and theorists can help us with. There is a wealth of work on racial injustice and gender injustice, their respective institutions of perpetration, and where they share commonalities and where they differ. Empirical investigations might, for example, reveal that institutions of policing form a key pillar of the race structure but are less central in the gender structure. This would suggest correlations produced by institutions of policing to be suspect with respect to discrimination on the basis of race, not sex.

The existence of multiple different hierarchical social structures also shows the approach of spelling out one-by-one an account of race, of gender, of class and adding the three together to be inadequate in formulating good theories of discrimination on the basis of race, gender, and class. Theorists of intersectionality point out that since multiple interlocking systems of oppression produce a form of oppression distinct from the sum of its parts, accounts of "single-axis" discrimination cannot capture all forms of discrimination. Thus, thick constructivism must be able to account for the fact that racial discrimination affects Black men and Black women differently. I do not yet have a fully adequate response to this important task, though I believe the challenge from intersectionality offers another reason why any plausible response to the puzzle of algorithmic discrimination must provide theories of those social categories on the basis of which agents discriminate. Thick constructivism allows for intersectional social categories such as *being a Black woman* to be defined by reference to the various ways that one's position in the hierarchical race structure as Black interacts with and modulates one's position in the hierarchical gender structure as a woman and vice versa. Those correlations produced by social factors that constitute the position of being a Black woman are potential enablers of algorithmic discrimination on the basis of the category *Black woman*. What exactly those social factors are, and how exactly they relate to those that constitute the category Black and the category woman, are key questions for a thick constructivist account of intersectionality. Here, I reemphasize the importance of social scientific study on how hierarchical structures interact to produce positions of multiple advantage, multiple disadvantage,

[35] Haslanger discusses her view of the relationship between social metaphysics and empirical social science in 'Race, Intersectionality, and Method: A Reply to Critics,' *Philosophical Studies* 171, no. 1 (2013), pp. 109-119.

and simultaneous advantage and disadvantage. In drawing on empirical work, thick constructivism tracks our changing awareness of how different structures of oppression and discrimination interact.[36]

4. Implications for Prevailing Theories of Discrimination

The thick constructivist account of race is a stark departure from conceptions of race implied in most work on discrimination. Most prevailing analyses assume a thin account of race defined by a set of phenotypic features that happen to carry social meaning. This might explain why past work has devoted notably little attention to developing an adequate *descriptive* theory of racial discrimination. But as argued in the preceding section, a thick constructivist account of race troubles our prevailing ideas of what it is to act on the basis of race, full stop—and as a result, challenges our prevailing assumptions and analyses of what racial discrimination even *is*.

One place where the standard machinery of discrimination theory seems to malfunction is the commonly held distinction between direct and indirect discrimination. While there is scholarly dispute about how to best characterize the distinction, the thought roughly is that indirect discrimination against Blacks harms Blacks by acting on some consideration that is not race "itself" yet still has negative disparate impact on the group. For instance, because Blacks have more often been subject to certain subordinating policing, they *indirectly* bear the brunt of policies that use past police encounters as a decision-making criterion. But if, as the thick constructivist holds, the category Black is defined in terms of those social facts that constitute its position in a hierarchical race structure, the gap between acting on "being Black" and acting on "having been targeted by violent policing practices" narrows significantly. If a given decision-making criterion has disparate impact on Blacks because it embodies a relation that constitutes Black qua thick social position, then acting on it does seem to be acting on "being Black" in an important sense—one that challenges the idea that doing so only acts on "being Black" indirectly.

More generally, without an accompanying theory of what race and racial categories are, the line between acting directly versus acting indirectly on the category can seem completely unfounded. This problem only becomes more acute when decisions are set by algorithmic tools such that references to mental states such as intentions and motives cannot distinguish direct from indirect discrimination. This is not to claim that there is no such distinction at all. In fact, the thick

[36] I thank a reviewer for pushing me to clarify how a thick constructivist solution handles the reality of multiply interacting axes of discrimination.

constructivist solution leaves room for a class of indirectly discriminatory actions: cases of acting on a factor correlated with but not constitutive of race. There is no deep incompatibility between the distinction and thick constructivism, but an account of that distinction must be *downstream* of a theory of race.

As noted in the last section, a thick constructivist interpretation of correlations with race does not propose a *particular* account of race that adjudicates what types of machine learning methods used on what kinds of data certifiably counts as discrimination on the basis of race. A thick constructivist account of race does not provide, say, an algorithmic solution to algorithmic discrimination. Nevertheless, statistical regularities with race disclose important causal consequences associated with the social position that those raced Black occupy. Non-accidental correlations between a given racial category and a set of social outcomes can reveal the social causal capacities that attach onto a given social position. A robust correlation between "Being Black" and "Being Picked Out For Stop-And-Frisk" contributes to our understanding of the causal consequences of being subject to norms and meanings that attach onto (supposed) morphological and ancestral characteristics commonly labeled "Black." Now comes the thick constructivist move: in observing more and more of these correlations with "Black," we might want to expand our account of "being Black" in terms of being socially positioned so to be probabilistically subject to a certain set of risks and pressures associated with being labeled "Black" (on the basis of facts about one's phenotype and presumed ancestry). These correlations might give a partial picture of what being raced Black is in the first instance.[37]

We can analogize this perspective to a similar effort in understanding class and class structure. Suppose I have a large dataset of individuals labeled with their "class category," defined in objectivist orthodox Marxist terms—your class is established entirely by your relationship to the means of production, so under capitalism you are either a member of the proletariat or the bourgeoisie—and I discover that being what Marx called a proletarian is correlated with having a long working day, limited personal property, and being subject to behavioral control on and off the job. I find that my algorithm can predict with great accuracy which of these individuals will

[37] Annabelle Lever's account of racism is similar to this thick constructivist account of race. She suggests anti-Black racism be conceived of as a broad set of harms suffered "as an ordinary part of daily life… simply because one is [B]lack rather than white." I suggest such "ordinary" experiences of racism may leave statistical markers that show us what constitutes the thick social position that is the racial category Black. I thank an anonymous reviewer for drawing my attention to this connection. 'Why Racial Profiling Is Hard to Justify: A Response to Risse and Zeckhauser,' *Philosophy & Public Affairs* 33, no. 1 (2005), pp. 94-110.

experience financial hardship in the next six months, and this result is heavily skewed toward those labeled "Proletariat" even if I remove the "Class Category" feature from my dataset. One interpretation of this phenomenon explains that being a member of this class causally influences or is causally influenced by having a long working day, limited personal property, and behavioral restrictions. This conclusion is fine enough, but there is something unsatisfying about the approach. When I use an algorithm to draw on the fact that you frequently experience financial hardship and behavior discipline, I act *directly* on the fact that you are in a *particular* precarious and subordinated position within the economic structure—a distinctive social position that perhaps warrants attention *as* a class category. This thought, however, is not available to me if I am pre-committed to an analysis of class that references only your position within the network of ownership relations.[38] If I hew strictly to the orthodox Marxian account, I only see those features as *causally related* to a class category rather than as themselves fitting together to *constitute* a class category. And depending on my particular theoretical and practical aims, it might help to adopt a theory of class that takes the latter route and defines class categories in terms of social factors that one is likely to be subject to having been deprived of access to the means of production. It seems right to me to say that a decision to refuse social services to an individual because she has a long working day, limited personal property, and is subject to behavioral control on and off her job discriminates on the basis of class—and this is a claim we can make on a thick constructivist conception of class. Thus, different accounts of class bear different theoretical and normative fruit. The same story applies to accounts of race.[39]

What fruit does the thick constructivist solution bear? As I have already argued, most important is its direct engagement with a theory of race needed to get the descriptive account of discrimination off the ground. Beyond this, thick constructivism also makes progress on debates regarding a *moralized* conception of discrimination. If racial categories are thick social positions in an unjust hierarchical race structure, then it is no surprise that actions that satisfy the descriptive analysis of discrimination on the basis of *being Black* are usually also acts of *wrongful* discrimination. These actions are likely to have many of the wrong-making features put forth by various theories of wrongful discrimination. It is easy to see how they might demean, compound injustice, emerge from

[38] Notice that my algorithm could even draw on your income or personal wealth, and on an orthodox Marxian definition of class, it would not be drawing on your class status.

[39] Nothing I say here is particularly novel to work on social ontology, see e.g., Aaron M. Griffith, 'Realizing Race,' *Philosophical Studies*, forthcoming.

racial animus, reinforce racial hierarchy, and socially subordinate.[40] Thick constructivism also suggests a natural answer to the question of which social categories and groups matter for wrongful discrimination: those categories constituted by social relations forming an unjust hierarchy, and especially those groups structurally disadvantaged as a matter of their position in that hierarchy.

Still it bears noting that thick constructivism and theories of discrimination are not typically fellow travelers in the philosophical and legal literatures.[41] Perhaps the thick constructivist account of race is too capacious or too activist-y to be plugged into a theory of racial discrimination. Indeed, the following thought seems to be a point of agreement among discrimination theorists: discrimination on the basis of race is a *particular* kind of racial wrong circumscribed within a broader set of racial wrongs. I tend to share this view. But it does suggest there to be something incompatible between thick constructivist accounts of race, defined in reference to an unjust race structure, and prevailing theories of discrimination on the basis of race. This presents a potential problem. Insofar as theories of discrimination on the basis of race (and gender and so on) must have a theory of what "race" and "gender" as social categories *are*, being incompatible with thick constructivism means needing to ally with some other school of thought—and in light of the algorithmic discrimination puzzle, theorists cannot continue to kick the can down the road on what accounts they *are* allied with.

Let me now venture into what I think is a predominant way of morally and legally reasoning about racial discrimination today: an analysis of discrimination that takes on board a causes interpretation of "because" and a thin account of race. To avoid an overly permissive view of what counts as an appropriate causal link with race, these efforts adopt an account that requires evidence

[40] For theories of wrongful discrimination based in: demeaning, see Deborah Hellman, *When Is Discrimination Wrong* (Cambridge: Harvard University Press, 2008); compounding injustice, see Deborah Hellman, 'Indirect Discrimination and the Duty to Avoid Compounding Injustice,' in Hugh Collins and Tarunabh Khaitan (eds.), *Foundations of Indirect Discrimination Law* (Oxford: Hart Publishing, 2018); racial animus and racist attitudes, see Richard Arneson, 'What is Wrongful Discrimination,' *San Diego Law Review* 43, (2006), p. 775; J. L. A. Garcia, 'Wrongful Racial Discrimination in Moral Analysis: Some Recent Accounts, an Alternative Conception, and Attempts to Extend Theoretical Models,' *Georgetown Journal of Law & Public Policy* 16, (2018), p. 697; reinforcing racial hierarchy, see Owen Fiss, 'Groups and the Equal Protection Clause,' *Philosophy and Public Affairs* 5, pp. 107-177; T. M. Scanlon, *Moral Dimensions: Permissibility, Meaning, and Blame* (Cambridge: The Belknap Press of Harvard University Press, 2010). Sophia Moreau presents a pluralistic account that covers multiple of these wrong-making features and highlights social subordination in *Faces of Inequality: A Theory of Wrongful Discrimination* (Oxford: Oxford University Press, 2020).

[41] One notable exception is Esa Díaz-León, 'On How To Achieve Reference to Covert Social Constructions,' *Studia Philosophica Estonica* 12, (2019), pp. 34-43, which discusses how discrimination on the basis of certain clusters of superficial natural properties played a part in the social construction of both the concept "race" and the social kinds that are "races." Díaz-León does not, however, apply a social constructivist metaphysics to an analysis of discrimination.

of a chain of *thin race*-disadvantaging to validate claims of racial discrimination.[42] In court battles, this exercise often cashes out as a statistical wild goose chase, with each side hiring their expert to perform all sorts of mathematical operations to show that the use of certain decision-making criteria does (or does not) discriminate against a certain racial group.[43] And since for them, racial categories refer to sets of phenotypic features and nothing more, statisticians can perform causal inference operations, cancelling out, so to speak, the effect of all sorts of *merely causally-related* attributes—socioeconomic status, neighborhood school quality, number of family members with a criminal history—to show that decision criteria exhibit no differential effects on race after all.

This statistical turn of reasoning about discrimination is unsurprising once we accept a thin account of race. It is also unsurprising that demanding statistical evidence of differential "race" effects, when the race concept is divorced from those social factors that constitute races qua social positions, presents an onerous burden of proof for those who claim discrimination—one that will only become even more burdensome when decision policies are set by algorithms. So it is no wonder that statistical efforts aiming to uncover some causal effect of thin race to prove racial discrimination often come up empty-handed. One interpretation of such outcomes is to concede that these methods show racial discrimination to be less pervasive than we might have thought—keeping in mind that it will be whites who will, due to their considerably rosier perceptions of racial equality, be more likely to accept the happy conclusion and racially oppressed groups who will overwhelmingly see it to be plainly false.[44] These stark differences show that basic facts regarding the prevalence and severity of racial discrimination as an empirical phenomenon are greatly contested.

[42] Accounts that define racial discrimination as differential treatment on the basis of a discriminator's perceived difference of race (qua superficial phenotypic features presumed to be linked to geographic ancestry) have a tough time making sense of "indirect" discrimination *as* discrimination. So, one way these accounts can find cases of disparate impact as bona fide cases of discrimination is via evidence of causal chains of thin race-disadvantaging, showing that thin race played a covert role after all as the basis of differential treatment. See e.g., Benjamin Eidelson, *Discrimination and Disrespect* (Oxford: Oxford University Press, 2015); Tarunabh Khaitan, 'Indirect Discrimination,' in ed. Kasper Lippert-Rasmussen, *The Routledge Handbook of the Ethics of Discrimination* (New York: Routledge, 2017).

[43] For example, even the plaintiffs in *Floyd v. City of New York*, a class-action lawsuit on whether the New York Police Department's stop-and-frisk policy stopped individuals on the basis of race, accepted that statistical evidence of stop-and-frisk rates in racial minority precincts had to be conditioned on what were implied to be non-race-related features such as crime in minority neighborhoods and housing vacancy rates to show the policy was discriminatory on the basis of race. Issa Kohler-Hausmann thoroughly chronicles the role of statistical reasoning about the causal effect of race in discrimination battles in 'Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination,' *Northwestern University Law Review* 113 (2018), p. 1163.

[44] Pew Research Center surveys on race show there to be significant gaps between responses by whites and those by racial minorities on virtually every single question about race relations, existing racial inequality, and racial progress. Excerpts from the 2019 survey: 84% of Blacks, compared to 54% of whites, say racial discrimination is a major reason Blacks have had a "hard time getting ahead." Large majorities of racial minorities (84% of Blacks, 67% of Hispanics, and 72% of Asians), compared to a slim minority of whites (48%), say that failure to note genuine cases of racial discrimination is a bigger problem than seeing racial discrimination where it does not exist. 35% of Blacks, compared to

Where do these divides leave us? For starters, as a matter of good political epistemology, I do not think we should accept the verdict that there is little racial discrimination. We have reason to reject analyses of race that stack the deck in favor of this conclusion and adopt a thick constructivist one instead. Here, political commitment and empirical consideration weigh in *substantively* on the upshot of any adequate theory of discrimination on the basis of race. But even more pressing is what persistent, massive discrepancies in perceptions of racial discrimination between whites and those who are racially oppressed might suggest about the dominant philosophical *method* of theorizing about discrimination and about race. There is no reason to believe that the premises from which we have often started our theorizing—that the paradigm cases of racial discrimination look like *this* and not that, that the concept should cover *these* kinds of cases and not these others, that judges and courts have ruled that *those* actions do not constitute bona fide discrimination—have not been formulated, implicitly or otherwise, under a dominant white perception of the rarity and particular distastefulness of racial discrimination. It is all too likely that most of our (carefully constructed) analyses are out of step with discrimination as understood and experienced by many racially oppressed persons. This is an unfortunate but wholly unsurprising way in which philosophical theorizing is itself marked by the social and material realities characteristic of particular raced positions. Here, I suggest thick constructivism to be well-suited to play another role in analyses of discrimination on the basis of race—as a lens through which we can respond to urgent but heretofore largely neglected second-order questions about philosophical theorizing on discrimination: whose 'discrimination' have we been theorizing? Whose *should* we theorize? To what extent does academic philosophy have the tools to do so? And finally, most importantly: what do we want out of our theorizing about racial discrimination? And can we do a *good* job without explicitly aiming to do a *just* job?

80% of whites, say Black people will eventually have equal rights in the U.S. 'Race in America,' *Pew Research Report*, Washington, D.C. (2019).