NYU | LAW

New York University School of Law

October 16th, 2025, from 4-7 pm
Lester Pollock Room, FH, 9th Floor

Colloquium in Legal, Political, and Social Philosophy

Conducted by
Liam Murphy and Samuel Scheffler

Speaker: **Deborah Hellman, University of Virginia**
Paper: **Why The Proxy Problem Is So Hard**

Colloquium Website: http://www.law.nyu.edu/node/22315

WHY THE PROXY PROBLEM IS SO HARD

October 1, 2025

Draft for NYU

Deborah Hellman (dhellman@law.virginia.edu)

Introduction: The Hard Proxy Problem

Computer scientists, legal scholars, and philosophers have been troubled for some time about what is referred to as "the proxy problem." The problem they have in mind is this. Legal and moral norms dictate that the use of certain attributes of individuals (their race and sex, paradigmatically) should only be used sparingly (if at all) in making predictions about or decisions that affect a person. One might assert, as courts sometimes do, that this prohibition rests on the fact that race and sex are irrelevant to whether a person will succeed at a job or repay a loan. Unfortunately, however, race, sex, and other protected attributes are not irrelevant in the following sense. They are useful in predicting many outcomes that banks, employers, and the state, are understandably interested in predicting, like whether an individual is likely to repay a loan, be a successful employee or recidivate. In other words, the law – and moral norms on which these legal prohibitions rest – restrict the use of race, sex, religion and other traits both when their use is irrational *and* when it is rational.

The fact that rational discrimination is legally prohibited (and rightly so, in my view) gives rise to a problem. Traits or features of a person or her situation that are connected in some way with protected attributes are often used to make predictions or decisions about individuals. In such cases, one might wonder whether laws and moral norms should also forbid the use of these related attributes on the grounds that they are "proxies" for race, sex, and other protected traits. Answering this question has become more pressing with the advent of machine learning tools. While banks, employers, prison officials and others who use machine learning [ML] algorithmic tools to make predictions about who will repay a loan, succeed on the job, and avoid recidivating ensure that these algorithmic tools are unaware of the race, sex, or other protected attributes of the people whose data on which these algorithms are developed and trained and of the individuals they evaluate or score, the algorithm detects other traits or combinations of traits that contain much of the predictive information of the protected trait. Because protected attributes are often correlated with (or predictive of) the outcomes of interest, machine learning systems will find other attributes that are themselves correlated with race, sex, etc. and use those features instead.

Some of the features that a ML algorithm will use may feel related or connected to the prohibited attribute in ways that seem morally or legally troubling. For example, a person's residential zip code or the fact that the person attended Wellesley College may be useful in predicting the outcome of interest because, at least in part, zip code correlates with race and attendance at Wellesley correlates with sex. If it is legally or morally problematic for a bank, employer or prison official to use a person's race or sex to make the relevant decision, is it also problematic for these actors to use zip code or attendance at a women's college to do so? And if so, is this because zip code is a *proxy* for race and having attended Wellesley is a *proxy* for sex?

The proxy problem has both a practical dimension and a conceptual dimension, as Gabrielle Johnson explains.[1] Practically, it is difficult to ensure that ML systems are not using attributes that may be proxies for protected attributes when making decisions or predictions. Were the law to forbid the use of zip code and attendance at a women's college by the bank or employer, the machine learning algorithm would uncover other traits or features that correlate with race and sex (perhaps slightly less well) instead. If those other attributes were then also prohibited, the algorithm would uncover others. This is not because the ML system is poorly motivated and "wants" to discriminate on the basis of race or sex. Indeed, such a system is not *motivated* at all. Rather, it is given a task (like predicting how likely a person is to repay a loan) and uncovers the features or combination of features that predicts that outcome. Even when the ML system has no direct information about the race or sex of the people it scores, it may well have information that while seemingly unrelated to loan repayment, turns out to be predictive of loan repayment because of some connection that this information has to the protected attributes. How can this process be arrested? This is the practical proxy problem.

To answer it, we first need to know what problem such a fix is looking to solve exactly. In other words, what makes a feature, or combination of features, a *proxy* for a protected attribute in the first place? Johnson calls this conceptual problem the "hard proxy problem," a term I borrow. The hard proxy problem is focused on "the theoretical relationship between proxy features and protected classes and asks when that relationship is meaningful."[2] To answer this question, we need a theory of proxies.

While the hard proxy problem has become salient in the AI context, it is not a parochial problem restricted to the domains of machine learning and AI. The conceptual question of what sort of connections make one trait a proxy for another (and especially for a protected attribute) is a live and important controversy facing policy makers in many areas. One salient example concerns university admissions. Following the Supreme Court's invalidation of the use of race in admissions,[3] university officials must consider whether it is legally permissible to adopt facially neutral policies if they do so, at least in part, because such a policy is likely to provide an admissions boost to underrepresented minority applicants. For example, suppose the university adopts a preference for first-generation [first-gen] college students in order to increase racial diversity. If so, is first-gen status a proxy for race?[4] And if so, what ensues from this designation, legally and morally?

[1] Gabrielle M. Johnson, *The Hard Proxy Problem: Proxies Aren't Intentional, They're Intentional*, Philosophical Studies (forthcoming).
[2] *Id*. at 3.
[3] Students for Fair Admissions, Inc. v. President and Fellows of Harvard College, 600 U.S. 181 (2023).
[4] The current federal executive branch appears to think so. A July memorandum from Attorney General Pam Bondi offers guidance to universities and other recipients of federal funds that "Facially neutral criterial (e.g. 'cultural competence,' 'lived experience,' geographic targeting) that function as proxies for protected characteristics violate federal law if designed or applied with the intention of advantaging or disadvantaging individuals based on protected characteristics." July 29, 2025 Memorandum for all Federal Agencies, "Guidance for Recipients of Federal Funding Regarding Unlawful Discrimination."

This Article aims to make progress on the hard proxy problem. In what follows, I defend two claims. First, I argue that "proxy" is a context dependent concept, by which I mean that defining a proxy for the discrimination context is importantly different from defining a proxy for other contexts like voting, for example. If this is correct, we are not going to make progress on the hard proxy problem by examining the concept of proxy in isolation. A theory of proxies for the discrimination context will depend on an account of what discrimination is. Second, I argue that there are several plausible conceptions of a proxy for the discrimination context if by plausible we mean that they reflect different ways that people use that term. At the same time, there is no way to choose among them, except by asking what the concept of a proxy for the discrimination context is *for*. Framing the hard proxy problem as a question of conceptual ethics reveals that the concept of a proxy for the discrimination context has a normative upshot. We are interested in proxies because we want to know what to do with them. To say that a neutral trait is a proxy for a suspect trait is to say that this neutral trait should be treated *as if* it were the suspect trait when evaluating the action at issue. If this is correct, solving the hard proxy problem will also require that one say something about why legal and moral norms proscribe (except in rare cases) the use of the protected attributes explicitly or directly. The hard proxy problem is hard therefore because its answer depends on the answers to these two difficult and contested questions.

In what follows, I will take up the hard proxy problem. Part I sets up the problem and introduces the terminology I will use to discuss it. Part II engages in a bit of conceptual ethics, arguing that we should focus on the normative proxy concept rather than the descriptive proxy concept for several reasons. Part III argues for the claim that an account of what makes one thing a proxy (in this normative sense) for another thing depends on the context in which this proxy concept operates and for this reason a theory of proxies for the discrimination context depends on what discrimination is and on why protected attributes are protected. Parts IV and V develop this argument, with Part IV focusing on the connection between an account of what discrimination is and a theory of proxies and Part V focusing on the connection between an account of why suspect traits are protected and a theory of proxies. Part VI makes this abstract discussion concrete by providing an illustration of how particular answers to each of these questions provides answers to some of the hypothetical cases discussed earlier in the article. A brief conclusion follows.

I. Conceptual Set-up

To begin, there are three possible ways that a person or system (including both a policymaker and an algorithm) might act in relation to a person with a particular trait S. First, the policymaker or algorithm might act on the basis of that trait or feature (S). Second, it might act on the basis of traits or features of the person other than S (N, which is not S). Third, it might act on the basis of a trait that is a proxy for S. On this way of cutting up the terrain, the trait that may be a proxy (P) is both distinct from S and also different in some way from N. If P is not distinct from S (and instead a constitutive part of S), then the actor acts on the basis of S. If P is not distinct from N, then perhaps there is nothing to the idea that some traits are "proxies" for others.

In what follows, I will use the following terminology.[5] Call P (for proxy) the trait or feature that might be a proxy. Call the trait that P may be a proxy for S (I use S because often this trait is a suspect or protected trait). And call T the feature that is of interest or that a ML algorithm is tasked to predict (the target). The question we are exploring then is this:

> *The Proxy Question*: When is P a *proxy* for S, in the context in which P is used (alone or together with other neutral traits) to select people with T.

Each type of action noted above (acting on the basis of S, acting on the basis of traits that are ~S, and acting on the basis of P (which while also ~S is distinct in some way such that it warrants different treatment) can lead to different legal treatment and moral assessment. When a person or entity acts on the basis of S and treats people with S differently than those without S, this constitutes disparate treatment on the basis of S.[6] Where S is a protected attribute, this action faces a high justificatory burden. Where S is not a protected attribute, it does not. The proxy problem with which we began was concerned with when and why certain features of a person or their situation are proxies for protected attributes like race and sex, so let's focus on the context in which S is the sort of trait whose use is either legally protected[7] or morally problematic. In the language of U.S. equal protection doctrine, we would say that S is a "suspect" or "quasi-suspect" trait.[8] If so, our second type of action is one in which an actor acts on the basis of a trait that is not S (or any other suspect trait). In such a case, the actor acts on the basis of a neutral trait or traits (N). Such actions are presumptively permissible (from the perspective of discrimination law and norms) unless these actions have a disparate impact on people with S. If they do create such a disparate impact, the actions face some justificatory burden but one that is considerably easier to pass than is required for disparate treatment on the basis of protected traits.[9] Third, an actor might act on the basis of P (a trait that is a proxy for S). In exploring the proxy relationship, then, we are interested in what connections make P a proxy for S and what the upshot of that determination yields in terms of how actions on the basis of P should be viewed or evaluated.

Now that we have both our terminology in hand and have sketched the possible ways that a person, institution or algorithm might act (on the basis of S, on the basis of ~S/N,

[5] Different authors writing about proxies use different terminology. *Compare* Michael Carl Tschantz, *What is Proxy Discrimination*, FAccT '22, June 21-24, 2022, Seoul, Republic of Korea, https://doi.org/10/1145/3531146.3533242.

[6] "Disparate treatment" is the term used by U.S. law. The law in other jurisdictions uses the term "direct discrimination."

[7] The trait could be legally protected by statute or, when the actor is a governmental entity, by constitutional law (in the United States).

[8] Equal protection doctrine treats race as a "suspect" trait and requires laws and policies that treat people differently on the basis of race to be subject to "strict scrutiny." Sex, by contrast, is a "quasi-suspect" trait and so subject to a lower justificatory burden termed "intermediate scrutiny." From here on, I will use the term "suspect" to refer to suspect traits and quasi-suspect traits, as both kinds must meet a higher justificatory burden than neutral traits and thus the question of whether other traits are proxies for these traits is relevant.

[9] The greatest difference between U.S. discrimination law and the discrimination law of other jurisdictions is that the difference between disparate treatment and disparate impact (or between direct and indirect discrimination) is less stark in the law of other countries. Citation.

and on the basis of P), we can articulate the two questions that our inquiry will focus on. First, what sort of *connection* between P and S or between P, S and T makes P a proxy for S? This is the *Proxy Question* articulated above. Keep in mind as we approach this question that it is possible that the connection between P and S is such that P is not a proxy for S but is instead constitutive of S (thereby blurring the line between acting on the basis of S and acting on the basis of P). In addition, it is also possible that there are no connections that establish a proxy relationship so that the answer to the proxy question is "never." If so, then all we have are two categories: acting on the basis of S and acting on the basis of ~S. In other words, we should not presume that there is such a thing as a proxy or that a proxy relationship exists. But if it does, there is a second question for us to explore. What is the significance of this connection between P and S (the proxy relationship) legally, morally, or in some other way?

It is worth emphasizing a point just mentioned: I am setting up the proxy problem by making the preliminary assumption that P is distinct from S rather than being a feature that is constitutive of T. That said, it is of course possible that a trait that we conclude is a proxy for S could also be constitutive of S. I will address this possibility in Part V.

The answers to the *Proxy Question* provided by scholars and courts to date can be fruitfully organized into three broad categories. First are those accounts that emphasize statistical or empirical connections. Second are those that emphasize explanatory connections. Third are those that emphasize intentional connections, in particular the intentions of the actor in using P to select for people with S.

The first family of views sees P as a proxy for S when P has a particular sort of statistical relationship with S and/or with T. This could be a simple statistical relationship, like that P is correlated (or strongly correlated) with S. Or it could be a more bespoke statistical connection. For example, consider the account provided by Anya Prince and Daniel Schwarcz.[10] According to Price & Schwarcz, P is a proxy for S when P's ability to predict T derives only from its correlation with the S. To put the point another way, if P would no longer be predictive of T if one controlled for S, then P a proxy for S.[11] What matters to Prince and Schwarcz is thus not the correlation between P and S on its own but the fact that this correlation is the reason (or the only reason) that P is predictive of T.

This approach has considerable intuitive appeal, which is illustrated with the following example drawn from Prince & Schwarcz:

> *Facebook Group*: The law prohibits life insurers from basing pricing and coverage decisions on genetic information.[12] A life insurer uses a ML algorithm to determine whether to offer life insurance, and at what rate, to each prospective purchaser. Because of the legal prohibition on the use of genetic

[10] Anya E. R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 Iowa L. Rev. 1257 (2020).

[11] *Id*. at 1262.

[12] *See* Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110-233 (2008) [GINA].

> information and despite the fact that genetic information is predictive of a person's likelihood of making a claim during the policy period, the algorithm used by the insurer does not use genetic testing results in its prediction model. However, the ML algorithm incorporates as an input whether a person is a member of a Facebook support group for people with a particular genetic disease.

In *Facebook Group*, according to Price & Schwarcz, membership in this Facebook group is a proxy (P) for genetic predisposition to disease (S) because membership in this Facebook group (P) predicts T (likelihood of making an insurance claim during the policy period) only because membership in the Facebook group (P) correlates with having a genetic predisposition to the disease (S).

This approach, which emphasizes the statistical connections between P, S, and T provides an intuitively appealing answer in this example – membership in the Facebook group does indeed feel like a *proxy* for the suspect feature, having a genetic predisposition to disease. However, the Prince & Schwarcz account fails to capture other examples which also, plausibly, feel like proxies.[13] In prior work, I use the example of unintentional redlining to illustrate the limitations of their account.

> *Unintentional Redlining*: A bank is legally prohibited from basing lending decisions on the race of prospective borrowers.[14] A bank uses an algorithm to predict which prospective borrowers will repay their loans. The algorithm was trained on data that lacked racial labels and is unaware of the race of prospective borrowers. The lender's algorithm uses the zip code of the borrower to predict loan repayment. As a result, Black and Latino borrowers are disproportionately excluded from access to loans.

In this example, zip code (P) would not be a proxy for race according to the Prince & Schwartz account because zip code is likely to be predictive of loan repayment, to a significant degree, even if one controlled for race. Or to put the point another way, zip code's predictive power is not entirely (or even mainly) due to its correlation with race. Rather, zip code is predictive of loan repayment because zip code not only correlates with race but also correlates with wealth and income, both of which predict loan repayment.

The second approach to identifying the sort of connections that establish the proxy relationship focuses on explanatory connections between P, S and T and in particular an explanation rooted in a history of discrimination or injustice. In prior work, I proposed (and then critiqued) one such account.[15] On that account, P is a proxy for S when P correlates with S and this correlation is explained by a history of discrimination or injustice.[16] On this explanatory theory of the proxy relationship, zip code (P) would be a proxy for race (S) in *Unintentional Redlining* because P

[13] I offered this critique of the Prince & Schwartz account in Deborah Hellman, *Defining Disparate Treatment: A Research Agenda for our Times*, 99 Indiana L. Rev. 205, 240 (2023).

[14] *See* Equal Credit Opportunity Act, 15 U.S.C. §§ 1691-1691f (2011) [ECOA].

[15] *See* Hellman, *Defining Disparate Treatment*, *supra* note 13.

[16] *Id*.

correlates with S and the reason that P correlates with S is (plausibly) due to a history of discrimination or injustice. Gabbrielle Johnson adopts a similar account. In her view, P is a proxy for S when "uses of that feature by the decision-maker to pick out the individuals that it does are in virtue of a causal-explanatory chain that is initiated by acts of discrimination directed toward members of those protected classes."[17] Importantly, on Johnson's view there need be no current correlation between the proxy trait and the suspect trait for a proxy relationship to exist.[18] For this reason, she terms her view the "Pure Explanatory Theory of Proxy Content."[19]

While these explanatory accounts capture plausible intuitions about whether zip code is a proxy for race in *Unintentional Redlining*, they fare less well in other cases. For example, they likely fail to capture *Facebook Group*. The reason that being a member of the Facebook group (P) picks out the people it does (T) is unlikely to be traced to discrimination or injustice. Even in a society without a history of discrimination against people with genetic predisposition to disease, people whose genes predispose them to disease, and their families and friends, may find comfort in associating and sharing experiences with others in the same situation.

A third account of the connections that create a proxy relationship is instantiated in U.S. constitutional law. On this view, intentions create proxies and without an intention to *use* P to select for S, P is not a proxy for S. The Supreme Court most clearly articulated this view in *Personnel Administrators v. Feeney* in which the Court considered whether a Massachusetts statute that provided a life-time preference for veterans in hiring for state civil service jobs should be treated as discrimination on the basis of sex.[20] The case was decided in 1979, a time at which women were excluded from most military service positions and 97.7% of veterans were men.[21] In *Feeney,* the Court held that unless the legislature had adopted the veterans preference "because of" and "not merely in spite" of its adverse impact on women, it would not be treated as disparate treatment on the basis of sex.[22] In our terms, the theory of the proxy relationship that underlies equal protection doctrine can be stated as follows: P (veteran status) is a proxy for S (sex) only if P is selected *because of* its correlation with S.

Just as with the two prior accounts of the proxy relationship, the intentional account also misses some cases that might plausibly be instances of proxies for suspect traits. On this view, neither *Facebook Group* nor *Unintentional Redlining* contain proxies. Indeed, on this view, there would be no (or almost no) proxies in the machine learning context. Unless a person deliberately directed a ML program to use the neutral trait to exclude people with a suspect trait, no proxy discrimination would occur.

Moreover, while U.S. constitutional law treats the intention to use P to select for people with S as a necessary element of the proxy relationship, existing law is unclear about whether such an

[17] Johnson, *supra* note 1 at 4.

[18] *Id*. at 10 (claiming that "the focus on statistical correlation, which is central to formal definitions of proxy use in machine learning and in theories of disparate impact law, is a red herring that can lead theories of discrimination astray").

[19] *Id*. at 1391.

[20] 442 U.S. 256 (1979).

[21] MAX CLELAND, 1979 ANNUAL REPORT: ADMINISTRATOR OF VETERANS' AFFAIRS 2 (1979).

[22] 442 U.S. at 279.

intention is sufficient to create a proxy. Indeed, this question is the site of the debate I mentioned earlier regarding whether facially neutral traits may be used deliberately to increase racial diversity.[23]

This brief summary divides the proxy accounts into three broad categories which I labeled statistical, explanatory and intentional. However, as this discussion has revealed but not yet highlighted, they are all *explanatory* accounts in some sense. For example, on the statistical account provided by Prince and Schwarcz, what matters is whether it is the statistical correlation between P and S that *explains* why P is predictive of T. On the so-called "Pure explanatory theory" offered by Johnson, what matters is whether a history of discrimination against people with S *explains* why P is predictive of T. Finally, on the intentional account found in U.S. constitutional law, what matters is whether the actor's intentions to use P to select for people with S explain why P predicts T. I began by calling the second account "explanatory" as that is the label Johnson uses for her theory. However, what is distinctive about it lies not in its focus on explanation but instead in the feature it identifies as what matters to that explanation: the history of discrimination that explains why P picks out the people it does. Going forward, therefore, I will refer to this type of account as "historical" rather than "explanatory."

II. A bit of conceptual ethics

Each of the proposed accounts of the proxy relationship seem to track a variant of how the term "proxy" is commonly used and each identifies a potentially important connection between P, S and T. So one might wonder, what criteria should we use to select among them? In the prior section, I used examples to suggest that each of the accounts of the proxy relationship got some examples right and some of them wrong. But I never specified what criteria should guide those intuitions (mine or yours). Was I asking you a descriptive question about how the term is currently used in a particular language community? Was I asking you an ontological question about what a proxy really *is*? Was I asking you a normative question about whether P should be treated as if it were S in the specific case at hand? Or, perhaps I should say, which question should I have been asking you?

Let me begin by noting that both the term and the concept of a proxy can be understood in both a descriptive and a prescriptive way. The term "proxy" is used to describe some sort of relationship between P and S (whatever that relationship consists in) and to say something about that relationship. In this way "proxy" is like other concepts related to discrimination each of which has both a descriptive and a moralized form. "Discrimination" itself is like this. Sometimes, when we say that A discriminates against B on the basis of S, we mean only to *describe* the fact that A treats B differently (or worse) than others because B has (or is believed to have) S. Such discrimination may be permissible or may be wrongful. Other times we mean to also say that this action is presumptively wrongful or legally suspect. For example, a law that

[23] The constitutional permissibility of doing so is the subject of much current debate. For arguments that it is permissible, *see e.g.* Sonja Starr, *The Magnet School Wars and the Future of Colorblindness*, 76 Stan. L. Rev. 161 (2024); Deborah Hellman, *Diversity by Facially Neutral Means*, 110 Virginia L. Rev. 1901 (2024). For an argument, that it is not permissible, *see* Brian T. Fitzpatrick, *Can Michigan Universities Use Proxies for Race After the Ban on Racial Preferences,* 13 Mich. J. Race & L. 277 (2007).

specifies that one must be sixteen years old to get a driver's license discriminates on the basis of age in the descriptive sense in that the law treats young people differently (and worse) than older people. However, it does not discriminate on the basis of age in the moralized sense because limiting driving privileges to those over 16 is clearly permissible. "Stereotype" and "bias" are similar concepts in this regard in that they can be used in a moralized or non-moralized manner.[24]

Embracing this plurality of usage (which I am not sure we could resist even if we wanted to), my first intervention would be to say that we need to pay more attention to whether we are talking about proxy as merely a descriptive relationship or instead as a prescriptive label that carries some normative upshot. Part of the confusion is this area can be traced to the fact that these meanings are often fused. The fact that a relationship between P and S exists that could plausibly be described as a proxy relationship does not show that the prescriptive meaning of proxy applies. For example, an adherent of the intentional view of the normative proxy relationship could concede that veteran status is a proxy for sex in the descriptive sense (due to the statistical correlation between veteran status and sex) while maintaining that veteran status is not a proxy for sex in the normative sense because the Massachusetts legislature did not adopt the preference for veterans in order to exclude women.

Second, I want to propose that the prescriptive sense of proxy and the proxy relationship is the more interesting and the one that we genuinely care about. And for these reasons, it should be the concept of proxy that we work to understand. In part, I think we should focus on the prescriptive conception of a proxy because I do not think there is any way to choose among the various descriptive accounts. Each describes a way that the term is actually used in the world and so all three of the accounts of what makes P a proxy for S seem like plausible accounts of the proxy relationship (in the non-moralized form). Proxy is a concept that we use to understand the world, rather than an entity or thing in the world that the concept should correspond to. As the term is used to describe multiple different sorts of relationships, it is hard to see what intuitions the examples used to argue for one account over another are trying to capture. Instead, to the extent the examples function as intuition pumps, the intuitions they pump are normative. In asking whether P is a proxy for S, the examples implicitly ask you to evaluate whether P should be treated *as if* it were S. Take *Facebook Group*. When you considered whether membership in the Facebook Group for people with a particular genetic disease is a proxy for having a genetic mutation that predisposes one to disease, what exactly were you asking yourself? I imagine that in answering that question, you implicitly asked yourself whether discrimination on the basis of membership in the Facebook group should be treated (by the law or morally) in the same way as would discrimination on the basis as a person's having a particular genetic mutation would be.

The first reason that the normative question is the one we ask ourselves, then, is traceable to the intractability of the question about which conception of proxy best captures its descriptive meaning. As a matter of how the term is used, all are plausible. In addition, there may be no answer to the question of what a (descriptively) truly is. For these reasons, I think we should

[24] For a non-moralized account of stereotype, *see* ERIN BEEGHLY, WHAT'S WRONG WITH STEREOTYPING (Oxford, 2025); for a moralized account *see* Lawrence Blum…. For a non-moralized account of bias, *see* Gabbrielle Johnson, …; for a moralized account, *see* THOMAS KELLY, BIAS: A PHILOSOPHICAL STUDY (Oxford, 2022)

accept that the descriptive conception of the term "proxy" captures several different and competing accounts.

The second reason we gravitate to the prescriptive question when confronting hypothetical cases is that it is this sense of the proxy that we *care* about when we talk about proxies. What we want to know is whether treating people differently on the basis of P should be treated *as if* it were on the basis of S. We care about understanding and defining proxies in the discrimination context so that we know what to do when an actor makes a determination on the basis of a trait that is plausibly a proxy for a protected attribute. If P is a proxy for a protected attribute, then treating people differently on the basis of P trait should be subject to the higher justificatory burden to which differentiation on the basis of the protected trait is subject. This proxy concept allows us to address the issues we care about in the context of discrimination. What we want to know is when P should be treated as if it were S, when S is a suspect trait.

That said, the moralized proxy concept we should focus on is one that is only weakly moralized – and so differs in this regard from the way that "discrimination," "bias," and "stereotype" function as moralized concepts. To see what I mean by this, contrast the weakly moralized proxy concept with both a non-moralized proxy concept and a strongly moralized proxy concept. The non-moralized proxy concept describes or captures a certain kind of connection between P and S (statistical, historical, intentional, etc.). By stating that P is a proxy for S (based on the connection or connections identified), this label has no moral valence and no normative upshot. By contrast, a strongly moralized proxy concept would have a normative valence. To say that P is a proxy for S would be to say something about the permissibility of an action that treats people differently on the basis of P. On the view I propose, the term *proxy* has a normative upshot but no moral valence. If P is a proxy for S, then P should be treated as if it were S when evaluating the action at issue. Because treating people differently on the basis of protected attributes faces high burdens of justification, such actions will often be impermissible but not always. On this way of understanding the proxy concept, use of a proxy for a protected attribute is not always illegal or morally wrong.

Let me summarize the conceptual clarifications and modifications I am proposing. The term "proxy," like several other terms related to the concept of discrimination (including "discrimination" itself, as well as "stereotype" and "bias") can be used in both a non-moralized and a moralized way. The non-moralized (or descriptive) sense of "proxy" captures (at least) three different sorts of relationships between P and S. Each of these capture a relationship between the purported proxy and the trait for which it may proxy. The statistical views emphasize the empirical connections – both simple and complex – between the group picked out by P and the group identified by S. The historical views emphasize that a history of discrimination or injustice against people with S explains the correlation between P and S or explains the connection between P and T. Third, intentional views identify a third potentially important connection between P and S, that is, the one created by the intentions of the actor who deliberately uses P to select for people with S. Each of these conceptions of a proxy seem to be plausible descriptive conceptions of what a proxy is and all are familiar (though Johnson's perhaps less so). I don't believe we have the resources to choose among them as understandings of what a proxy is, if that term is understood in its non-moralized sense.

What we care about when we care about proxies is not this descriptive sense of the proxy relationship but instead a moralized conception of a proxy. We want to know when treating people differently on the basis of P should be treated as if it were on the basis of S instead. The interesting proxy concept is the one with normative upshot.

I am not proposing that we abandon the non-moralized conception of a proxy. But I hope we will be more attentive to the fact that the term is used in both a moralized and non-moralized sense. In addition, I propose that we accept that all three accounts of the what the proxy relationship (descriptively) is – in the non-moralized sense – are plausible views and each tracks ways in which the term is used. What we must take care to avoid is the unsupported assumption that because one of these relationships exist, then P is a proxy for S in the moralized sense. For that claim, one needs an argument that justifies why P should be treated as if it were S. In the next section, I make some progress in answering this question.

Before doing so, I want to briefly address the objection that I too quickly reject the ability to identify which *descriptive* conception of the proxy relationship is the correct one and, relatedly, that I fail to appreciate the importance of doing so. While I believe that we care most about the normative proxy relationship, it is surely true that isolating the correct/best understanding of the descriptive proxy concept could also be useful and interesting, if it is possible to do so. In saying that what we care about when we care about proxies is the normative proxy concept, I should not be understood to say that the descriptive concept is uninteresting. My claim is more modest: the descriptive proxy concept is *less* interesting.

Moreover, I argued that we lack the criteria to select one descriptive proxy concept as the correct one. Gabrielle Johnson thinks otherwise and offers a theory of proxy content that she takes to flesh out the *descriptive* understanding of the proxy concept. Yet in building her argument for the particular theory she adopts, she specifically draws on the reader's *normative* intuitions about specific cases. If the examples feel "morally dubious," then, in her view, they are more likely to deploy proxies for suspect attributes. She specifically defends this approach in the following way: "The methodological assumption is that if a case feels morally dubious, the proxy use in that case must have a significant, explanatorily robust connection to the protected classes. While not infallible, this heuristic is practically useful since our intuitions are strongest in such cases."[25] In other words, though she sets out to identify when a purported proxy trait is "meaningfully about" the suspect trait, she uses moral intuitions about when that use is problematic to tell her when this is the case. While she believes this method does not make her proxy account one that identifies a normative proxy, her method is unable to assure us that that is the case. If the intuitions on which it draws are about what is morally problematic, which likely relate to intuitions about when P should be treated as if it were S, then she may actually be tracking a normative proxy relationship.

[25] Johnson, *supra* note ___ at 13.

Of course, it is possible that Johnson has a reply or someone else could explain why one of the plausible descriptive accounts of the proxy relationship is the best one and the only one we should use. Until then, however, I think we should recognize that there are multiple plausible accounts of proxy as a descriptive concept and no criteria by which to adjudicate among them. It is this lack of criteria that is crucial. For while, as we will see, there are also multiple contenders for the normative proxy concept as well, and robust disagreement about which is best, theories about what discrimination *is* and *why* protected traits are protected provide the criteria to select among them.

III. Proxy as a context-dependent concept

Each of the accounts of the proxy relationship described in the prior section put forward an account of what makes P a proxy for S as if that account applied in all contexts in which the proxy concept is used. At the same time, the examples that each author used in their discussion were set in the context of a question about discrimination. Does that matter? Are we investigating a generic concept of proxy or instead something that might better be referred to as proxy$_{\text{discrimination}}$. And if proxy$_{\text{discrimination}}$ [herein after "proxy$_{\text{D}}$"] is different from proxy$_{\text{something else}}$ [herein after "proxy$_{\text{X}}$"], then the concept of a proxy may depend in a critical way on what discrimination is.

Is proxy$_{\text{D}}$ different from proxy$_{\text{X}}$? To investigate that question, let's examine proxies in a different context.[26] Consider proxy voting. While laws governing voting in political elections in the United States,[27] do not permit voting by proxy, proxy voting is permitted in other countries[28] and contexts, including voting by corporate shareholders and voting by members of private clubs or associations. [29]

Consider the following hypothetical example:

> *Proxy Voting*: Sarah is a member of a club that permits club members to vote on club policies by proxy if they are unable to attend club meetings. In order to do so, the club rules specify that a member designate another club member as their proxy and sign a form giving this person written permission to exercise

[26] While I will argue that the appropriate concept of a proxy for one context is different than for another, the two are not so different that they fail to refer to the same more general concept. A proxy *stands in for* the thing for which it is a proxy in each of the cases discussed. In that sense, proxy$_{\text{discrimination}}$ and proxy$_{\text{voting}}$ are not homophones, like bank (where you put your money) and bank (the edge of the river).

[27] *See Election Crimes and Security*, FED. BUREAU OF INVESTIGATIONS, https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-frauds-and-scams/election-crimes-and-security (last visited Feb. 18, 2025) (defining "[v]oting more than once or using someone else's name to vote" as "[f]raud by the [v]oter" and a federal election offense); Andrew Tutt, *Choosing Representatives by Proxy Voting*, 116 COLUM. L. REV. SIDEBAR 61 (2016) (explaining that proxy voting is "to this author's knowledge, . . . not used in any American jurisdiction").

[28] Thomas Heinmaa, *Special Voting Arrangements (SVAs) in Europe: In-Country Postal, Early, Mobile and Proxy Arrangements in Individual Countries*, INT'L INST. FOR DEMOCRACY AND ELECTORAL ASSISTANCE (Oct. 19, 2020), https://www.idea.int/news/special-voting-arrangements-svas-europe-country-postal-early-mobile-and-proxy-arrangements (showing that proxy voting is not allowed in many Eurasian countries, including Denmark, Germany, Ireland, Russia, and Spain).

[29] *See 17* C.F.R. § 240.14a-4 (2021); *see, e.g.,* VA. CODE ANN. § 55.1-1953 (describing the proxy voting process for homeowners' associations).

> their vote. Sarah selects Paul as her proxy and signs the appropriate form. Paul is therefore a proxy for Sarah.

Note how different this understanding of the connection that establishes the proxy relationship is than some of those canvased in Part I. Some of those views required a correlation between the two.[30] For example, Prince & Schwarcz focused, in part, on how much overlap there was between the group defined by membership in the Facebook group for people with a particular genetic mutation (P) and the group with the particular genetic mutation (S). In the context of voting by proxy, this concern seems inapt. Indeed, it would be hard to imagine how it would even apply. Instead, it is the designation, along with the satisfaction of the writing requirement, that appears to create the proxy.

By contrast, the intentions-based view of the connection that creates a proxy relationship might seem more in line with the conception of proxy we see in the voting context. But that observation is too quick as intentions play a very different role in creating a proxy in the voting context than they do for the intentional theory of proxies in the discrimination context. In particular, the accounts differ with respect to *whose* intentions are relevant in each context. In the voting context, it is the intention of person for whom the proxy will stand in that is relevant, not the intentions of some other actor. We can see this by considering the following case:

> *Deviant Voting Rules*: Suppose a club policy states that if a member (M_1) wishes to vote by proxy, the Club President can identify any other club member that the Club President chooses (M_2) to act as the proxy for M_1 and M_1 can have no authority over how M_2 votes. In such a case, club rules specify M_2 should be treated as M_1.

The club rules specify that M_2 is a proxy for M_1 in this context and so provide a descriptive account of the proxy relationship for the Club. Yet, there is an important sense in which the person picked by the Club President would not be a proxy for M_1 in the normative sense. M_2 is chosen by the Club President. While the Club President may have selected M_2 because the President believes that M_2 shares the policy preferences of M_1, it is also possible that the President chose M_2 because M_2 shares the President's own policy preferences.[31] Given what voting is, M_2 is not a voting proxy for $M_{1,}$ in the normative sense.

What this contrast between $\text{proxy}_{\text{voting}}$ and $\text{proxy}_{\text{discrimination}}$ illustrates is that the normative proxy [NP, for short, going forward] concept is importantly dependent on the context in which it operates. Who counts as a NP for the voting context depends on what voting is and its normative foundation.

An account of a NP that is appropriate for the discrimination context also depends on that context. In particular, a NP for the discrimination context depends on what discrimination is and

[30] *See also* Tschantz, *supra* note __ at __.

[31] Whether a purported proxy selected by the Club President because that person shares the policy preferences of the club member for whom they would be a proxy is an appropriate voting proxy or a deviant voting proxy is a more difficult case and may depend on whether one thinks voting is normatively grounded in autonomy or as a means of vindicating the interests or preferences of voters.

on why discrimination law and norms treat some traits as protected, and their use suspect, and not others. To determine whether some P is a NP for some S then requires answering both of these questions. Because the answers to each of these questions are contested, it is no surprise that the hard proxy problem is so hard. In the next two sections, I fill in this account. Part IV demonstrates the way that different accounts of what discrimination *is* determine a threshold that limits which proxy accounts are possible accounts of a NP for the discrimination context. Part V then turns to the answers to the normative question: why are protected traits protected? It shows how different answers to this question supplement the threshold account of Parv IV, providing a full account of a NP for the discrimination context.

IV. The connection between what discrimination is and the normative proxy relationship

Recall, "discrimination" is a term that can be used in a moralized and non-moralized fashion. In this section, I am focusing on what discrimination is, in the non-moralized sense, such that an action can be an instance of discrimination without also being wrong. Kasper Lippert-Rasmussen defines discrimination in the following way: "to discriminate against someone is to treat her disadvantageously relative to others because she has or is believed to have some particular feature that those others do not have."[32] This formulation is fairly uncontroversial, but what is controversial is what "because" means in that definition.[33] Lily Hu helpfully parses two contrasting views of the meaning of "because" here as the *reasons* view and the *causes* view.[34] On the *reasons* view, "because" is understood to refer to the discriminator's own reasons for acting. A discriminates against B on the basis of some feature S if A *takes S as a reason* for treating B differently than she treats another. On this view, it is A's own reasons that matter in the sense of the reasons that were reasons for her. Hu uses the term "operative reasons" to refer to reasons in this sense. These are reasons from the point of view of the agent whose reasons they are.[35]

On the *causes* account, by contrast, "because" is understood in terms of the factors that caused A to treat B differently. The case of unconscious bias illustrates the contrast between these two accounts.

> *Unconscious bias*: Jack is considering hiring two job candidates for one position: Jane and Jamal. Jane is white and Jamal is black. Jack consciously rejects race as a relevant factor to making this decision. However,

[32] Kasper Lippert-Rasmussen, *Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination* (Oxford: Oxford University Press, 2014) at 15.

[33] The Supreme Court case *Bostock v. Clayton County*, 590 U.S. __ (2020), focused on precisely this question when ruling on whether Title VII of the Civil Rights Act of 1964, as amended, which prohibits discrimination "because of" sex also prohibits discrimination because of sexual orientation or gender identity and found that it did. The Court's understanding of "because of" as "but for" causation sparked much controversy and critique. *See e.g.* Mitchell N. Berman & Guha Krishnamurthi, *Bostock Was Bogus: Textualism, Pluralism, and Title VII*, 97 Notre Dame L. Rev. 67 (2021).

[34] Hu, *supra* note 40.

[35] Some authors use the term "motivating reasons" to refer to such reasons, which seems felicitous to me, but this term has led to confusion because some authors include unconscious motivations within that term. What distinguishes "reasons" from "causes" in the way I am using this term is whether the agent him or herself takes the reasons to be a reason herself.

> unconsciously Jack perceives Jamal as less qualified due to his race. As a result, Jack hires Jane.

On the *reasons* account, Jack does not discriminate against Jamal *because of* his race because Jamal's race was not the fact that Jack himself took as a reason for downgrading the strength of Jamal's application. Rather, Jack's reason was something else, even if his perception of that something else (qualifications, experience, etc.) is subtly and unconsciously affected by Jamal's race. As a result, *Unconscious Bias* is not discrimination on the basis of race, on the *reasons* account. On the *causes* account, by contrast, Jack does discriminate against Jamal *because of* his race as Jamal's race caused Jack to downgrade the strength of Jamal's application.

Whether one understands discrimination in terms of reasons or instead in terms of causes has implications for what concept of a *normative proxy* is appropriate for the context of proxy discrimination. If discrimination is limited to the situation in which an actor herself takes S as *her* reason for acting, then cases of proxy discrimination will be limited to cases in which the actor deliberately adopts P in order to select for (or against) people with S. As a result, there will be no instances of unintentional proxy discrimination.

On this understanding of a NP (which depends on a *reasons*-based understanding of discrimination), ML algorithmic systems will only very rarely deploy proxies for protected traits. As the ML system itself has no genuine reasons for its actions in the sense of reasons that it takes to bear on how to act, there will be no proxies. Unless a human being instructed the system to find and use proxies for protected traits, there will be no cases of proxy discrimination that occur with ML algorithms.

If you, the reader, do think that membership in the Facebook group in *Facebook Group* is a NP for genetic information, there are two ways around this conclusion. First, you could reject the *reasons* view of discrimination and instead adopt the *causes* view, discussed below. Alternatively, you could argue that certain processes that the ML system adopts are relevantly like acting for a reason such that they should be treated as the same thing. Whether some algorithmic processes *are* relevantly like a person acting for a reason is a big question that I cannot address here. That said, I highlight this avenue in case the reader is tempted to explore it.

The reasons view of what discrimination is thus creates a fairly demanding threshold. Indeed, of the three types of views we canvassed earlier only the intentional proxy view passes that threshold and is thereby a possible accounts of a NP for the discrimination context. A *causes* account of discrimination is much more permissive, establishing a threshold to be sure but one that each of the families of views we discussed can easily pass. If causes are what matter to assessing whether discrimination on the basis of S occurs, then accounts that posit a causal connection (of some sort) will be plausible accounts of a NP for the discrimination context.

To see the model, return to the case of *Unconscious Bias* in which Jack's perception of Jamal's race causes Jack to downgrade Jamal's application. Race is a cause of Jack's decision and so on the *causes* view of what discrimination is, Jack discriminates on the basis of race in downgrading Jamal's application. In proxy cases, that causal relationship is mediated by the proxy. In other words, the protected trait (S) is causally related to the purported proxy (P) which is causally related to the outcome of interest (T).

Take, for example, the situation in which an algorithm uses income to predict loan repayment. In this case, it is plausible to think that race is causally related to income either directly (because of racial discrimination in employment) or indirectly (because race causes poor educational opportunity, which causes a particular level of income). If race causes income (to some degree) which causes T (likelihood of repaying the loan), then the causes account of discrimination would seem to be amenable to the view income is a proxy for race because race is causally related to the purportedly discriminatory action. These causal connections are likely to encompass the various statistical relations (correlation alone, or correlation plus something else) and so the statistical accounts of the proxy relationship pass this causal threshold.

Next consider the intentional proxy account. Suppose an actor selects the neutral trait in order to pick out people with the suspect trait. For example, a university adopts an admissions preference for first-gen students in order to increase racial diversity. Race causes first-gen status (in part), which causes admission (in part). In so saying, we might be simply describing a causal pathway that is similar to the one described above. A person's race may be causally connected to the fact that no one in her family has graduated college. But on that account of the relevant causal pathway, the intentions of the actor play no role. Alternatively, race may cause first-gen status in a different way: the fact that first-gen students are disproportionately members of racial minority groups is the reason (at least in part) that the university adopted the preference for first-gen students. The intentional theory of proxies thus also passes the threshold established by the causal account of what discrimination is.

Finally, consider the historical account. On this account, income is proxy for race because the reason that income picks out the people it does is due to a prior history of race discrimination. In other words, race causes income which causes the selection of the particular people selected (T), at least in part. This is a similar causal pathway as the first account. As a result, the historical theory of proxies also passes the threshold of the causal theory of discrimination.

So far, we've looked only at the descriptive question regarding what discrimination *is*. The answer to this question provides a threshold condition for a NP. What we have seen is that the reasons account provides a demanding threshold: because discrimination relates to the reasons for which a discriminator acts, only those features that were selected in order to pick out people with the protected attributes are potential normative proxies. The causes view, by contrast, is more permissive. On this view, each of the theories of proxies we canvassed and that are in common usage and discussed in the literature describe plausible accounts of a normative proxy. (That is not to say that the causes view is entirely permissive however. Features that are not causally connected along any pathway to the protected attribute are ruled out as proxies.)

The prior section argued that defining the normative proxy concept depends in important ways on the context in which it operates. When we are talking about proxies for the discrimination context, therefore, the normative proxy concept depends on what discrimination *is*. Because the nature of discrimination is itself disputed (between the *reasons* view and the *causes* view), it is unsurprising that people disagree about which understanding of a normative proxy is best. The intentional view aligns with the *reasons* understanding of discrimination. On the *causes* view, by contrast, all three understandings of what makes P a NP for T are plausible.

But assessing whether each proxy account meshes with what discrimination *is* is only a first step. If some P is a normative proxy for some S, then P will be treated as if it were S. To determine when P *should be treated* as if it were S, we need to know why some S's are treated as morally and legally suspect in the first instance. A normative proxy is defined not only by reference to the answer to the question about what discrimination *is*, but also by reference to the answer one provides to the question of why protected traits are protected. I turn to that question in the next section.

V. The connection between a theory of suspect traits and the normative proxy relationship

Defining a normative proxy for the discrimination context depends on an account of why P should be treated as if it were S in that context. To see why, return to the voting context and the example of *Deviant Voting Rules*. In *Deviant Voting Rules*, the Club rules specified that the Club President could choose any other club member to act as the proxy for the member who was unable to attend. These Club rules have force, and so in some sense they do create a proxy relationship (in the descriptive sense) because they establish that M_2 should be treated as M_1 when the Club President so designates. But, in another sense, this rule is open to critique when the concept of a proxy is used in the moralized sense – in the sense that P should (morally-speaking) be treated as if it were S. Given what voting is and its normative foundation in autonomy, M_2 should not be treated as if she were M_1 because a person's vote should be an exercise of autonomy. So M_2 is not a normative proxy in fact.

The reason that M_2 is not a normative proxy in the *Deviant Voting Rules* rests on a view about the normative underpinning of voting. Similarly, in order to define a normative proxy for the discrimination context, we must ask when it is that P should be treated as if it were S given the normative foundations of discrimination law (or norms). Answering that question will depend on why suspect traits are protected within such an account. The hard proxy problem is hard therefore not only because it depends on whether discrimination should be understood in the reasons sense or the causes sense, but also because it depends on why protected traits are protected. Both questions are contested and difficult, and it is thus unsurprising to find that people disagree about when and why some trait is a proxy (in the normative sense) for another.

Unfortunately, answering this second question is bigger and more contested than the dispute between the reasons and causes conceptions about what discrimination *is*. It is therefore harder to straightforwardly map out different normative accounts about why protected traits are protected (which will naturally relate to their underlying views about what makes discrimination wrong) and explain how they each would bear on different possible accounts of normative proxies. That said, this article has made progress on the hard proxy problem by *locating* the disputes that make the hard proxy problem difficult to resolve and the relevant criteria that one would bring to bear to settle these disputes. In order to determine when some P is a proxy (in the normative sense) for some protected attribute S, we must resolve two questions. First, is discrimination best understood in the *reasons* sense or the *causes* sense? Second, given the normative foundation of antidiscrimination law and norms, why are protected attributes treated differently than other traits? Answers to these questions fill out the conception of a normative proxy.

Interestingly, this account of normative proxies relieves some pressure on another contested question: how to define protected attributes themselves. Race, sex and other such traits have uncertain and contested boundaries.[36] And this difficulty in defining legally protected attributes is becoming increasingly salient in legal disputes about questions of discrimination.[37] Consider a few recent examples. *Bostock v. Clayton County* considered whether Title VII"s prohibition on discrimination on the basis of sex encompasses discrimination on the basis of sexual orientation and gender identity.[38] While the Court in *Bostock* held that discrimination on the basis of sexual orientation and gender identity was also prohibited by the statute, it did not do so on the grounds that the protected attribute "sex" includes sexual orientation and gender identity, as was suggested by some amici.[39] Yet, other cases addressing whether prohibitions on transwomen or transmen using bathrooms or participating in sports teams that align with their gender identity may well force this definitional question. The boundaries of the category race are similarly being contested. *Students for Fair Admissions v. President and Fellows of Harvard College*, which disallowed the use of race as a preference in university admissions, explicitly noted that a university could permissibly take into account the ways in which race had affected a particular student's experience.[40] In so doing, the opinion created a puzzle about when and why a decision is on the basis of race in the forbidden sense versus in the permissible sense.[41] Third, *Haaland v. Brackeen*[42] considered whether a preference for adoptive parents who are members of the same or different tribes found in the Indian Child Welfare Act is a racial preference. While the case was dismissed on standing grounds, the issue is likely to recur.

These are difficult questions. Hu calls attention to this contestation regarding the metaphysics of protected categories in the context of debates about whether various ML algorithms discriminate on the basis of race.[43] According to Hu, whether an algorithm discriminates on the basis of *race* or instead on the basis of a "factor *correlated with race*" depends on the understanding of race one adopts. In Hu's view, this "metaphysical" question must be answered *before* one can ascertain whether an algorithmic system discriminates on the basis of race.

[36] *See e.g.* Joshua Glasgow, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer, What is Race?: Four Philosophical Views (2019); [Fill in with example of literature contesting sex] – maybe Lily and Issa's paper.

[37] *See* Deborah Hellman, *Defining Disparate Treatment: A Research Agenda for Our Times*, 99 Indiana L. J. 205 (2023) (describing four puzzles that defining disparate treatment involves, two of which are focused on defining protected traits).

[38] 590 U.S. 644 (2020).

[39] Brief of Philosophy Professors as *Amici Curiae* in Support of the Employees.

[40] 600 U.S. 181 (2023).

[41] For a detailed examination of that puzzle, *see* Benjamin Eidelson and Deborah Hellman, *Unreflective Disequilibrium: Race-Conscious Admissions After* SFFA, *American Journal of Law and Equality* (2024) 4:295-325.

[42] 599 U.S. 255 (2023).

[43] Lily Hu, *What is 'Race' in Algorithmic Discrimination on the Basis of Race*, Journal of Moral Philosophy.

Hu herself adopts a "*thick* constructivist" account of race that "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*."[44] But is the thick constructivist account of race and other suspect traits correct? Not everyone agrees with this position. Happily, if the upshot of determining that a feature is a normative proxy for a protected attribute is that we should treat discrimination on the basis of the proxy *as if* it were discrimination on the basis of the protected trait, in cases where some trait may be either a normative proxy or a constitutive part of the protected trait, it will not matter practically which we conclude as both should be treated as discrimination on the basis of the protected attribute.

To illustrate, consider the following example:

> *Wheel-chair user*: An employer refuses to hire anyone who uses a wheelchair.

In *Wheel-chair user*, is wheelchair use (P) *a proxy* for disability (S) or is wheelchair use an *aspect of* disability?[45]

If discrimination against wheelchair users is (at least) a normative proxy for disability, then it will not matter whether we conclude that wheelchair use is a constitutive aspect of disability.

We will be unlikely to avoid all contested questions about the scope of protected attribute as some features that a thick constructivist account of race, sex, or other protected attributes will find to be constitutive of these attributes may fail to count as normative proxies (given the particular definition in use). But we are likely to greatly diminish the significance of the problem of defining protected attributes by turning to a normative theory of proxies. I take this to be a salutary upshot of this account.

Part VI: An application

While answering the two questions about discrimination (what it is and why protected attributes are protected) is beyond the scope of what it is possible to do in this paper, let me close by illustrating how my own views about the answers to these questions illuminate the examples thus far discussed. Doing so will show how answers to these questions about what discrimination is and why protected traits are protected yields distinctive answers to the hard proxy problem.

In my view, A discriminates against B on the basis of S when A's perception of S *causes* her to treat B differently than she would have otherwise. In other words, my answer to the first question about what discrimination is adopts the *causes* account rather than the *reasons* account. Second, in my view, protected traits are protected because of their expressive potential. Race, sex and other suspect attributes are different from unprotected traits (like the letter that starts one's last name, for example) because treating people differently on the basis of race and sex is

[44] Hu, *supra* note 40 at 3.

[45] An appellate court used precisely this example as one "'proxy' situation[] where a case can be made for 'constructive' disparate treatment, if not actual disparate treatment." *See* McWright v. Alexander, 982 F.2d 222, 228 (1992).

more likely to express denigration and so to be demeaning than is an action that treats people differently on the basis of an unprotected trait. I will not defend these views here.[46] Instead, I will briefly explain what account of normative proxy they give rise to and how that account would resolve some of the examples we have discussed.

First, because I reject the *reasons* view of what discrimination is and instead adopt the *causes* account, the fact that one trait is deliberately used in order to pick out people with a different trait has so special significance. For this reason, ML systems can employ proxies. In addition, when such an intention is present, this fact does not yield the conclusion that the purported proxy is a normative proxy. It might be (because the reason is a cause) but it might not. Whether it is will depend on whether the purported proxy is a trait that has a similar expressive resonance as the protected attribute.

To illustrate how this view plays out, consider the following examples. The first example relates to the issue regarding college admissions after *SFFA*.

> *Diversity by Facially Neutral Means*: The law forbids colleges and universities from counting the race of an applicant as a plus in college admissions. Suppose a college awards an admissions preference for applicants who would be first-generation college graduates if they are admitted and successfully graduate. Further, suppose the reason the college does so is in order to admit more Black and Latino applicants than would be admitted without this preference.

In *Diversity by Facially Neutral Means* [*Diversity*], the fact that first-gen status is selected in order to increase racial diversity has no special relevance because while intentions are one causal pathway by which the prohibited attribute may yield the result at issue, this causal pathway is not privileged. For this reason, first-gen status may be a proxy for race, but it may not. Whether it is will depend on whether first-gen status has a similar expressive potential as does race itself. In my view, it does not. Use of race to make admissions decisions is much more likely to express denigration of the people selected than is first-gen status. For this reason, first-gen status should not be treated as if it were race in *Diversity* and so first-gen status is not a normative proxy here. The fact that many people would use the word "proxy" to describe what is going on in *Diversity* does not count against this account because the word has multiple descriptive meanings, one of which focuses on intentional connections. The fact that first-gen status is a proxy for race in the descriptive sense does not settle the question of whether it is a proxy in the normative sense.

Compare this case with the Wellesley College example. In that example, an ML system downgraded the job applications of graduates of two women's colleges, Wellesley and Smith. Graduation from Wellesley is not intentionally deployed in order to exclude women but this fact has no special relevance. Sex is causally related to attendance at Wellesley, which in turn is causally related to the downgrading of a person's application. That relationship is sufficient to pass the first hurdle of the normative proxy relationship. However, attendance at Wellesley is not a normative proxy for sex unless attendance at Wellesley also has a similar expressive

[46] I develop and defend this account in Deborah Hellman, When is Discrimination Wrong? (Harvard Press, 2008).

resonance as does sex. Here I think that what these two features express is very similar. Wellesley is commonly understood to be a women's college and so excluding graduates of Wellesley expresses much the same thing as would the exclusion of women. For this reason, attendance at Wellesley is, in my view, a normative proxy for sex.

The case of *Unintentional Redlining* is also plausibly a normative proxy but is a much closer case than the Wellesley example. In *Unintentional Redlining*, race does cause the excluded borrowers to be excluded (at least in part) because race is causally related to zip code either directly or because race is causally related to wealth and income which are causally related to zip code. But, like in the two prior cases, zip code is not a normative proxy unless zip code has a similar expressive resonance as does race itself. Here I think the case for a normative proxy is stronger than in the first-gen example but weaker than in the Wellesley example. While zip code is a very different trait than race, lending exclusions on the basis of zip code have a history that is tied to race. It is the expressive associations this history generates that make the use of zip code that make zip code a plausible normative proxy for race. Note, however, that it isn't the history of discrimination itself that does the work. For this reason, the view I am putting forward is meaningfully different from Johnson's historical account of proxies. Rather, it is the social salience of that history that matters because that social salience changes the meaning of what use of this feature expresses.

This section illustrates how one account of what discrimination is and why suspect traits are protected will fill in the account of normative proxies. I cannot argue for my substantive views about these questions here, nor do I have space to consider more examples or objections to these brief assessments of individual examples. Nonetheless, I hope it provides the reader with some sense of how a theory of discrimination relates to a theory of proxies.

Conclusion

This article has made progress on the hard proxy problem in several ways. First, I have distinguished the descriptive and the prescriptive sense of the proxy relationship. Second, I have argued that each of the common accounts of the proxy relationship are plausible accounts of the descriptive conception of proxy and that the moralized proxy concept should be understood as weakly moralized, so that it carries the normative upshot that P should be treated as if it were S. Third, I have argued that this moralized proxy concept is dependent on the context in which it is used in the following way. A normative proxy for the context of discrimination depends on how one understands what discrimination is and on why protected traits are protected. Answering these two questions will yield distinct answers to the normative proxy question of when P should be treated as if it were S in the action at issue. The hard proxy problem is thus especially hard, because it is interwoven with other complex and contested questions. While the upshot of this analysis is that there are no easy answers available, it is nonetheless illuminating as it allows us to see the interrelationships between debates about proxies and debates about what discrimination is and what makes it wrong.