ANTIDISCRIMINATORY PRIVACY

Ignacio N. Cofone*

Draft 2017 for PRG. Please ask me for an updated version before distributing further. Thank you!

Abstract

The paper examines the information dynamics of privacy and discrimination to design anti-discriminatory privacy rules, especially for statistical and algorithmic discrimination. To do so, it uses empirical studies of informational anti-discriminatory rules and explores how privacy rules can overcome the limitations that these rules faced.

It proposes that taste-based discrimination and statistical discrimination, a traditional distinction in economics, have the same information dynamic and should therefore be addressed similarly by privacy law. The common element between different kinds of discrimination is that, to effectively prevent them, informational rules must focus on blocking information flows that can be used to shift discrimination to other groups (e.g. former inmates versus black men). Anti-discriminatory privacy rules, in other words, should block not only undesirable information but also their proxies.

The paper develops a theory on how to identify such proxies based on the cross-elasticity of information. It then applies this idea to algorithmic discrimination and proposes that the literature has so far brought legal solutions to an information problem. The paper proposes an information solution to the informational problem instead.

TABLE OF CONTENTS

Table of Contents..........................................................1

I. Introduction.....................................................................2

II. Privacy versus Antidiscrimination ................................4

* Research Fellow, NYU Information Law Institute. Contact: inc2@nyu.edu.
A. Bias-based and statistical discrimination ........................................ 4
B. When privacy fosters discrimination ............................................. 7
C. Case studies ................................................................................. 9

III. Discrimination as an Information Flow Problem .................. 11
    A. A typology on when to use privacy rules ............................... 11
    B. The role of proxies .................................................................. 13
    C. Cross elasticity ......................................................................... 14

IV. General Implications ............................................................... 18
    A. Equal protection ....................................................................... 18
    B. Bias-based versus statistical discrimination distinction ......... 22
    C. Scope of the proposal ............................................................. 23

V. An Application: Algorithmic Discrimination ....................... 24
    A. Unbiased decision-makers ...................................................... 24
    B. How algorithms discriminate ................................................. 28
    C. Privacy solutions to algorithmic discrimination ................. 30

VI. Conclusion .................................................................................. 32

I. INTRODUCTION

It is widely recognized that “[a]lthough litigation is important for bringing to light purportedly objective requirements perpetuating workforce segregation, prevention is key to eliminating systemic discrimination.”1 In this paper, I offer a novel way to achieve the aim of preventing discrimination before it can take place. I propose that, next to the standard efforts of antidiscrimination law to regulate behavior by curbing how people use of information about others to install discriminatory practices, there are benefits to

1 MARIE MERCAT-BRUNS, DISCRIMINATION AT WORK: COMPARING EUROPEAN, FRENCH, AND AMERICAN LAW 108 (2016) (arguing that there is an interest in institutional changes that focus “on mechanisms of inclusion over causes of exclusion: exploring other measures inciting people to take preventive action against the causes of discrimination or to establish institution-wide safeguards”).
regulating the *acquisition* of such information as well. Concretely, I explore the information dynamics of privacy and discrimination to design anti-discriminatory privacy rules.

This argument challenges the universal applicability of the belief that privacy and antidiscrimination are always opposing, sometimes found in social sciences, as well as the belief that they are always complementary. To challenge both these views, I use two canonical case studies that are often used to illustrate the privacy versus discrimination dynamic. The first is a well-known study conducted on auditions by female musicians to reduce gender discrimination in orchestra hires. The second is the attempt to protect people who have been in prison by banning “the box” in employer forms that ask whether the candidate has a criminal record.

These case studies are useful to explore how privacy rules can overcome their limitations for antidiscrimination aims. I propose that, while bias-based, statistical, and algorithmic discrimination work differently in many ways, they all have the same underlying information dynamic. The crucial task for preventing discrimination through privacy rules is identifying and blocking data points that are believed to be accurate proxies for the category that the law wants to protect. Anti-discriminatory privacy rules must focus on blocking information flows that can be used as proxies to shift discrimination to other groups.

The caveat to this principle is that, for bias-based discrimination, anti-discriminatory privacy rules must be conscious of their detrimental long-term effects, usually but not always offset by their short-term gains. There is a tradeoff between assimilation and avoiding the need for assimilation. For statistical and algorithmic discrimination, this problem is not present.

For the three kinds of discrimination, privacy rules can offer short-term protection from discrimination by blocking the data point that the law considers harmful to take into account in a decision-making process. Privacy rules can aid, in this way, anti-discriminatory efforts.

This lesson is particularly useful to address algorithmic discrimination, which I use as an application for the framework
proposed. The key insight that anti-discriminatory privacy brings to algorithmic discrimination is that, while most of the approaches so far have brought legal solutions to an information problem, an information solution to the information problem can address its root more effectively.

[signposting paragraph]

II. PRIVACY VERSUS ANTIDISCRIMINATION

A. Bias-based and statistical discrimination

Broadly speaking, there are two reasons why a decision-maker might engage in discrimination. The first occurs when the decision-maker has an irrational “taste” for discrimination: he is a bigot or has animus against a particular group of people. We could say that he is biased against such group. I will call this bias-based discrimination. The second occurs when the decision-maker uses someone’s gender, race, sexual orientation, or any other characteristic, to make a statistical inference. It doesn’t assume animus or bias. This is called statistical discrimination.

The standard economic narrative is that a large amount, if not most, of discrimination is of the statistical type.

---

4 This is often referred to as taste-based or animus-based discrimination; I will call it bias-based to emphasize its cause and avoid confusing it with a voluntary act.
This idea can also be found in legal scholarship. It is a well-established idea in economics that, when a decision-maker (Abby) wants to learn information about another individual (Ben) and such information is impossible or costly to acquire, the decision-maker will rely on statistical generalizations of the groups that the individual belongs to. This is commonly referred to as statistical discrimination.

In Phelp’s seminal statistical discrimination model, an employer cannot observe workers’ level of skill $q$, drawn from a normal skill distribution, but he can observe their group identity, and a noisy productivity signal. We can define such identity in any way, such as $P$ for purple hair and $G$ for green hair. Under the model, the question of statistical discrimination is the question of why two workers with the same productivity signal, but from different groups, are treated differently.

This can take place under two conditions. The first condition takes place when the groups’ signals are equally informative but the employer believes that one group, $P$, has a lower average human capital investments leading to lower average skill. It does not matter whether the identity is probative of skill level, but only whether the employer believes so. In that case, employers will consider employees from $P$ to have a lower expected productivity, so $P$ workers will receive a lower salary under the same signal. The second condition takes place when the skill distributions are identical but the signals for $P$ workers’ skills are less informative than those of $G$ workers. Because this will mean that the expected productivity

---

10 Phelps, supra note 6.
11 $N(\mu_j, \sigma_j^2)$
12 $\theta$, where $\theta = q + \varepsilon$, $q$ being skill and $\varepsilon$ being a normally distributed zero-mean error.
13 $j \in \{P, G\}$
of a $P$ worker with any signal will be closer to that of the population average, this will lead to highly qualified $P$ workers to receive a lower salary than their $G$ equivalents, and low qualified $P$ workers to receive a higher salary than their $G$ equivalents.\textsuperscript{14}

Twenty years ago, we started to see empirical evidence of this process. In a field experiment involving fictitious automobile buyers, Ayres and Siegelman showed that car dealers quoted significantly lower prices to white men than to female and black buyers.\textsuperscript{15} In a groundbreaking field experiment ten years later focusing on race, Bertrand and Mullainathan showed that resumes with white-sounding names have fifty percent more chances of receiving a callback than identical resumes with black-sounding names.\textsuperscript{16}

Title VII of the Civil Rights Act bans using this type of statistical generalizations to exclude some historically disadvantaged groups on a set of dimensions, such as race and gender. This is the case even when the generalization is true.\textsuperscript{17} These empirical studies raised critical questions about the effectiveness of title VII to prevent discrimination in the workplace.\textsuperscript{18} Antidiscrimination law, some argued, is

\begin{footnotesize}
\begin{enumerate}
\item Hamming and Moro, \textit{supra} note 7 at 137–140.
\item Marianne Bertrand & Sendhil Mullainathan, \textit{Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination}, 94 AM. ECON. REV. 991–1013 (2004) (showing that, for employers in Boston and Chicago, resumes with white-sounding names have fifty percent more chances of receiving a callback)).
\item City of Los Angeles Dep’t of Water and Power v. Manhart, 435 U.S. 702 (1978).
\item Angela Onwuachi-Willig & Mario L. Barnes, \textit{By Any Other Name: On Being Regarded as Black, and Why Title VII Should Apply Even If Lakisha and Jamal Are White}, 2005 WIS. LAW REV. 1283–1344, 1284 (2005).
\end{enumerate}
\end{footnotesize}
ineffective at dealing with unconscious biases.\textsuperscript{19} This leads to asking what alternative and complementary methods exist to combat discrimination.

\textbf{B. When privacy fosters discrimination}

Oftentimes in social sciences, and particularly in economics, discrimination is described as a problem of not having enough information about others.\textsuperscript{20} According to this account, having insufficient information leads people to recur to heuristics to judge others, which can easily result in false opinions. These false opinions, in turn, are attributed to everyone who falls under the heuristic, resulting in beliefs that could be racist, sexist, or homophobic.\textsuperscript{21}

Making information about oneself more available would therefore avoid the need for such heuristics and reduce discrimination, while having more privacy would worsen it.\textsuperscript{22} Representing the standard economic account, Strahilevitz has argued that \textquote{by increasing the availability of information about individuals. We can reduce decisionmakers\textquotesingle reliance on information about groups} and that, therefore, \textquote{there is often an essential conflict between information privacy protections and antidiscrimination principles, such that reducing privacy protections will reduce the prevalence of distasteful statistical discrimination.}\textsuperscript{23}

The underlying idea of this account is that, in many contexts, society must tolerate statistical discrimination because the only way to dispense it is to provide the decision-maker with more information, and there is a normative reason

\begin{itemize}
\item \textsuperscript{19} Lu\textsuperscript{\textquoteleft}in Wang, \textit{Race as Proxy: Situational Racism and Self-Fulfilling Stereotypes}, 53 DEPAUL LAW REV. 1013–1110, 1017 (2004).
\item \textsuperscript{20} E.g., Lior Jacob Strahilevitz, \textit{Privacy versus Antidiscrimination}, 75 UNIV. CHIC. LAW REV. 363–381 (2008).
\item \textsuperscript{21} See Id.
\item \textsuperscript{22} See Id.
\item \textsuperscript{23} Id. at 364.
\item \textsuperscript{24} Id. at 364. See also Lior Jacob Strahilevitz, \textit{Reputation Nation: Law in an Era of Ubiquitous Personal Information}, 102 NORTHWEST. UNIV. LAW REV. 1667–1738, 1682–1688 (2008).
\end{itemize}
for which he should not have access to it. Increasing privacy, therefore, could only make matters worse.

This idea permeates interactions outside of the academic discourse as well. For example, the LGBTQ movement implicitly used this rhetoric to advocate coming out as a political strategy. Their idea was that increased visibility would lead the general population to see that they interact with LGBTQ people every day, and to a decline in prejudices. Their understanding views privacy and discrimination as existing in tension.

Some legal scholars, however, have argued for specific institutional contexts that limiting information can aid in antidiscriminatory efforts, particularly for landlord-tenant relationships, genetic information, and disability law. Roberts, in particular, has claimed more generally that privacy and anti-discrimination are complementary.

The universal applicability of both beliefs—that privacy and antidiscrimination are always opposing or always complementary values—is questionable. Privacy rules, I suggest here, have an appropriate scope in anti-discriminatory efforts. To identify such scope, we must ask how and when privacy rules can be used as a tool to fight discriminatory decisions.

In the next section, I introduce two canonical case studies. The first is a well-known study conducted on auditions by female musicians to reduce gender discrimination in symphony orchestra auditions, and it has been used to show privacy’s compatibility with antidiscrimination. The second is

25 Strahilevitz, supra note 25 at 1723–1736.
28 Id. at 2103–2127. (focusing on disability law but also extrapolating those conclusions to make a general argument).
29 Id. at 2103–2127. (clarifying that this is despite the different normative values held by privacy law and antidiscrimination law).
30 Robert Post, Prejudicial Appearances: The Logic of American Antidiscrimination Law The Brennan Center Symposium on
the attempt to help people who have been in prison reintegrate to society by banning “the box” in employer forms that ask whether the candidate has a criminal record, and it has been used to show privacy’s tension with antidiscrimination.\(^{31}\) I use both of them to show in which cases privacy rules are compatible with antidiscrimination efforts, and in which cases they are not.

C. Case studies

In 2000, a group of economists conducted a well-known study that provided a novel way to test for sex-biased hiring on symphony orchestra auditions.\(^{32}\) Most of the previous economic literature on discrimination had focused on disparities in earnings,\(^ {33}\) and few were able to address actual hiring practices.\(^ {34}\) Even though orchestra directors are the kind of well-trained professionals that one expects not to have gender biases, and even though symphony orchestras had a fairly transparent hiring procedure, a gap between the proportion of female elite music school graduates and female elite orchestra hires prevailed.

After auditions were held behind a physical screen (oftentimes a curtain), preventing those hosting auditions to know auditioners’ gender, female hires increased by one third.\(^ {35}\) The screens increased the ex-ante probability of each woman to pass the initial round by 50%.\(^ {36}\) The study generated a research context where discrimination in the workplace seemed pervasive.

---

\(^{31}\) Strahilevitz, supra note 21.


\(^{33}\) See e.g. BECKER, supra note 3.

\(^{34}\) Goldin and Rouse, supra note 33 at 715.

\(^{35}\) Id. at 716. (stating that “the screen increases the probability a woman will be advanced out of a preliminary round when there is no semifinal round”).

\(^{36}\) Id. at 738.
Another field experiment that speaks about this issue studies the “ban the box” policy. The “ban the box” policy prohibits employers to ask applicants before interview stages whether they have a criminal record (a box often found in application forms with yes and no checkboxes). It is an attempt to protect people who have been in prison and helping them re-insert themselves into society, given that having a criminal record is a substantial barrier for obtaining employment.\(^\text{37}\) In many states, the initiative also explicitly aimed to protect black men to the extent that they disproportionately have criminal records, so they should disproportionately benefit from equaling employment possibilities for people who hold those records.\(^\text{38}\)

By sending fictitious job applications to entry-level positions in states that implemented the policy, Agan and Starr showed that “ban the box” backfired by fostering statistical discrimination based on race. \(^\text{39}\) This result had been anticipated before,\(^\text{40}\) and it had been argued that “empirical estimates indicate that employers who perform criminal background checks are more likely to hire black applicants


\(^{39}\) AGAN AND STARR, \textit{supra} note 39.

than employers that do not,” but the effect remained to be proved experimentally.

Agan and Starr showed that, in the absence of the policy, white applicants received 7% more callbacks than similarly qualified black applicants but, after the policy went into effect, the gap increased to 45%. This increase in the gap by a factor of more than six means that the policy was a loss for black applicants without a criminal record and a win for white applicants with one. “[W]hen employers lack individualized information about criminal history, they tend to statistically generalize that black applicants are likely to have records and white applicants are likely not to have them,” although the effect exaggerates actual racial differences in criminal records.

III. Discrimination as an Information Flow Problem

A. A typology on when to use privacy rules

As it can be seen in the case studies, blocking an information flow did very well for combating discrimination in one case, and not so well in the other. One could be tempted not to see this interaction as involving a privacy rule because it does not refer to the larger societal values that privacy often does, such as autonomy or personhood. The reason why one can and should call it a privacy rule is that the information dynamic is the same as that of privacy: a channel of information flow is deemed undesirable and is blocked. While the social goals here might be different than in most cases in which privacy is established, the interaction between two people is intervened in a similar vein. Both the screen in the symphony orchestra study and the “ban the box” policy in the field experiment produced a rule that blocked an information

---

41 Holzer, Raphael, and Stoll, supra note 8 at 472.
42 AGAN AND STARR, supra note 39.
43 Id. at 25, 29–30.
44 Id. at 4.
45 Id. at 26.
46 Roberts, supra note 28.
flow. By doing so, the rule forced a decision-maker to choose without a particular data point that the policy did not want taken into account.\textsuperscript{47}

In any case, whether we call it privacy is not central to adopting the tool that I propose here. One could read this paper as “anti-discriminatory information rules” and the central idea would be unchanged. But becoming conscious of this informational parallel allows one to see a new dynamic between privacy and antidiscrimination. Privacy has long been known to safeguard wide societal interests. For example, Post has said that privacy attempts to protect rules of civility, and the identities of individuals in a community.\textsuperscript{48} Privacy has also been said to protect reputational interests.\textsuperscript{49} Privacy can be similarly used to protect society’s antidiscrimination interests as well.

The question that emanates from the case studies, and from the discussion outlined above, is when is privacy in tension with antidiscrimination, and when is privacy an effective tool to combat discrimination. After seeing that privacy has, in some cases, aided in anti-discriminatory efforts, and seeing how it worsened discrimination in others, it is of vital importance to determine under which conditions blocking an information flow will achieve this aim.

This leads us to a typology. There are two conditions under which privacy rules can reduce discrimination. First, it is desirable to block future information when information samples are expected to be skewed, when only a non-infinite amount of information can be gathered. For example, if the New York Police Department decided to engage in predictive policing based only on prior arrest numbers, and it had more of its police force in Bronx than in other boroughs, it would be likely to make more arrests there than anywhere else, leading

\begin{footnotesize}
\textsuperscript{49} Ignacio Cofone & Adriana Robertson, \textit{The Privacy Bell}, (2017).
\end{footnotesize}
to more comparative police presence in Bronx, leading to more comparative arrests, and so on and so forth, independently of the actual crime rates. More information would be unhelpful for the New York Police Department in that case, and they would be better off if that information flow were blocked.

Second, people have limited time and attention to receive information, even when the information is unlimited.\textsuperscript{50} Moreover, once people form a belief, most do not update their priors as cleanly as an ideal rational actor would when new information arrives.\textsuperscript{51} The importance of this effect will be different depending on context, and on the informational demands of the decision: simple decisions that require little information are likely to be updated better by new information than complex decisions that are either rooted in deep beliefs or producing an information overload.\textsuperscript{52}

The key element is noting that, when a data point is blocked with the intention of preventing a discriminatory decision, oftentimes there are proxies available for such blocked information.

\textbf{B. The role of proxies}

The considerations set above allow one to examine the set of conditions under which privacy will foster discrimination, and the ones under which it will prevent it. Why did a privacy rule work to stop discrimination against female musicians, but backfired for ban the box?

The difference between one case and the other in terms of information is the existence of proxies that allowed the decision-makers to gauge the data that the rule was trying to block. Orchestra directors could not know through any other information point whether each, say, violin player was male or female (they might have guessed had it been an opera audition). Employers in states that implemented the “ban the box” policy, on the other hand, knew or believed that black

\begin{flushright}
\begin{tabular}{c}
50 \\
51 \\
52
\end{tabular}
\end{flushright}
males had a higher probability of having a criminal record than other applicants.

For ban the box, privacy rules were less effective not only—or even mainly—because they failed at protecting the group that they intended to protect, but because they shifted the existing discrimination from a disfavored group to proxies that affected another group disproportionately. This was especially damaging because the second group (black males) was historically a socially disfavored and worthy of protection as well—for many, even more than people with criminal records.\(^{53}\)

Shifting discrimination in such way is only possible when proxies to the blocked piece of information are available—especially when they are believed to be accurate and are easily observable. Those proxies can be used by the decision-maker with equally negative or even more devastating results than the original discriminatory decision. Therefore, it is crucial for any effective antidiscrimination-motivated privacy rule to identify and block those information flows as well. Checking for the existence of these proxies is necessary to predict the effectiveness of any anti-discriminatory privacy rules and of crucial importance to design better ones.

This analysis would have helped predict that “ban the box” would not achieve its regulatory aims. It also illustrates that blocking those proxies as well could have reversed the unintended effect that it had. For example, the regulation could have mandated the use of initials (and not full names) in resumes, making it more difficult to identify race before the interview.\(^{54}\) This would have effectively impeded gender and race based discrimination until the interview stage by eliminating available proxies, achieving the policy’s aim.

C. Cross elasticity

\(^{53}\) See Strahilevitz, supra note 25 (arguing that the case shows that some types of statistical discrimination must be tolerated).

\(^{54}\) Bertrand and Mullainathan, supra note 17 at See generally supra (showing that names work as accurate predictors of race, and that people with black-sounding names receive fewer call-backs than people with white-sounding names).
A crucial task, then, is to identify which pieces of information are proxies for others. A first approximation could lead one to measure this objectively, and to consider that proxies are about correlation between two data points. Under this conception, if Ben does not know $X$ about Abby, but he knows $Y$, and $Y$ happens to correlate with $X$, then $Y$ could be seen as a proxy for $X$. However, it is not the objective relationship between $X$ and $Y$, but their relationship in Ben’s mind, which will determine the extent to which he is likely to use $X$ as a proxy for $Y$. Therefore, whether a data point qualifies as a proxy for another data point for our purposes, is a question of cross elasticity.

[define cross elasticity]

Let $A$ be Abby, $B$ Ben, $X$ the unavailable information that Ben would like and $Y$ the information that Ben could use instead. Let $c^X$ be the cost for Ben of acquiring information $X$ about Abby, and $c^Y$ the cost for him to acquire information $Y$ about her. Let $q$ be quantity. Let $c_1$ and $q_1$ refer to the cost and quantity of information at time 1, and $c_2$ and $q_2$ refer to the time and quantity and time 2. When the cost of acquiring $X$ increases, the proxy relationship between information $X$ and information $Y$ is given by the change in the quantity of $Y$ that Ben attempts to acquire divided by the change in cost of the now more costly piece of information $X$.\footnote{Like with any other measure of elasticity, elasticity cannot be measured when there is no change in cost.}

$$E = \frac{c^X_1 + c^X_2}{q^Y_1 + q^Y_2} \frac{\Delta q^Y}{\Delta c^X}$$

If the elasticity between the data points is negative, this denotes that they are complementary pieces of information, while if it is positive, it denotes that they are substitutes.\footnote{$E \in (-1,1)$.} A piece of information is a proxy for another one for Ben if he treats them as substitutes.

An example might clarify this. Imagine Abby applies for jobs with Ben, Carlos, Dylan, and Eric. Imagine Ben, Carlos, Dylan, and Eric only care about productivity, and they want to know whether Abby is likely to ask for a maternity leave.
during the duration of the employment (X). Imagine state laws protect Abby from having them ask her this directly.

Before the law, Ben, Carlos, Dylan, and Eric could just ask Abby this directly, which would take them about 2 minutes long between their question and her answer. Take each minute to cost them 1 imaginary utility point. After the law, they need to ask five indirect questions instead of one to get that same information, taking him five times as long during their brief interview and therefore taking them 10 minutes instead. So state law increased their acquisition cost of information X by 500%.\(^57\) They can now choose between incurring this increased cost, or spending the amount of time for one question (2 minutes) asking a different one: whether Abby is engaged or recently married (Y). Imagine Ben still asks Abby about her potential maternity leave through five indirect questions, and that Carlos, Dylan, and Eric inquire about marital status instead.

The price of X changed from 2 to 10, and the quantity of X demanded changed from 4 to 1. Therefore, the cross elasticity of X and Y is

\[
E = \frac{2 + 10}{4 + 1} \times \frac{3}{8} = 0.9
\]

In this example, employers treated marital status (on average) very much as a proxy for likelihood of asking for a maternity leave. An elasticity of 1 would have meant that they

\(^{57}\) The cost of X will be determined by its Shannon entropy \(H(p)\), where \(p\) is the probability distribution of \(X\) and \(H(p) \in (0,1)\). A variable's entropy is a measure of its uncertainty. This is, of how much information, on average, one needs to know its value. \(X\) has a higher entropy than \(Y\) when we need more yes/no questions to know its value. See THOMAS M. COVER & JOY A. THOMAS, ELEMENTS OF INFORMATION THEORY 12–15 (2012). See generally Claude Shannon, A Mathematical Theory of Communication, 27 BELL SYST. TECH. J. 379 (1948).

Since, for these purposes, we care about elasticity, the relative entropy of \(X\) and \(Y\), which is the distance between their probability distributions, is also relevant as it will determine cost differences. The relative entropy of \(X\) and \(Y\), \(D = (p \parallel z)\), will describe the inefficiency of describing \(X\) (with a probability mass function \(p\)) through \(Y\) (with a probability mass function \(z\)). See COVER AND THOMAS, supra note at 18–21.
treated them as perfect substitutes, and an elasticity of 0 would have meant that they treated them as completely independent. Given the elasticity of 0.9, if state law really wants to protect Abby, it should prohibit employers from asking her about her marital status as well.

Alternatively, imagine that Ben hires a Human Resources company to interview job candidates. The company offers two services: since it’s forbidden to ask the candidates whether they want to have kids, a basic service asks them whether they recently married, while a more complete service estimates their likelihood of getting pregnant directly through a series of estimators with almost complete accuracy. How much is Ben willing to pay for the second service compared to the first? If he is not willing to pay much for the difference, we can infer that he considers the data points to be substitutes.

The relationship between these data points can be identified in a similar way when the law makes X not more difficult but impossible to acquire. In this situation costs and prices are absent. After the state law is enacted, Ben, Carlos, Dylan, and Eric are unable to incur a higher cost to acquire X like Ben did in the previous example and ask Abby about a potential maternity leave through a larger set of questions. In this new scenario, imagine that, when asking about future pregnancy was allowed, 75% of employers do so (Ben, Carlos, and Dylan), and 25% did not (Eric). When the law forbids it, 50% ask about marital status (Ben and Carlos), and 50% do not (Dylan and Eric). The degree of substitution of the data points is given by

\[ E = \frac{\Delta Q^Y}{\Delta Q^X} \]

In our example,

\[ E = \left| \frac{0 - 50}{75 - 0} \right| = 0.66 \]

As it can be seen, determining whether a data point is a proxy for another in terms of antidiscrimination is an empirical question, but not one of objective correlation but one of subjective perception measurable through cross elasticity. A sound empirical strategy to interrogate whether a concrete data point functions as a proxy, therefore, would not use
quantitative analysis to measure predictiveness but a field experiment or survey to measure perceptions, and evaluate cross elasticities based on those.

IV. GENERAL IMPLICATIONS

A. Equal protection

The struggle over the understanding of equal protection since *Brown v. Board of Education* has been articulated as falling under two competing conceptions of equal protection: anticlassification and antisubordination. The conventional view is that courts predominantly take an anticlassification position, limiting the scope of equal protection to treating all persons with equal civility and respect, but refraining from using equal protection as a tool to redistribute, accommodate, or object to disparate impact. Siegel’s historical account argues that American law has shifted ambivalently from one to the other.

A natural question that arises in this context is whether the tool proposed here operates under the logic of antisubordination or that of anticlassification. The answer is that it operates under both. One does not need a notion of antisubordination to block a flow of information (which, in some way, operates as color blindness), which makes this

---

60 Cite doctrine. See also Washington v. Davis, 426 US 229 (1976) (on disparate impact, taking an anticlassification approach).
61 Siegel, * supra* note 56. Siegel also traces a third understanding, in between these two, under which equal protection strives not to achieve color blindness or protection from subordinating practices, but protection from the threat of society's balkanization. See Reva Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE LAW J. 1278–1366 (2011) (arguing that courts concerned with antibalkanization focus on diversity more than on equality).
proposal viable under current anticlassification-dominated equal protection doctrine even outside of the scope of Title VII. At the same time, in virtue of being only an option that can operate alongside standard antidiscrimination law mechanisms, and not the way to address discrimination, the tool proposed is fully compatible with the logic of antisubordination.

Moreover, by focusing on the acquisition of suspected information rather than on its use, and by depriving decision-makers of these information points, anti-discriminatory privacy avoids the need for proving what is in the mind of the employer.62 This is a significant advantage since this proof is difficult to obtain as employers hide discriminatory intent, especially when there are mixed motives.63

While being operable under an anti-classification paradigm is an advantage for this tool in the sense that it makes it compatible with the mainstream of antidiscrimination doctrine and case law, this also implies its central limitation. While anti-discriminatory privacy rules work well to avoid disparate treatment, it has a limited effectiveness when dealing with some aspects of disparate impact. Most notably, anti-discriminatory privacy rules would not be useful to address cases in which affirmative action is considered desirable.64

A possibility to make this tool compatible with affirmative action or other concerns for diversity, albeit rarely applicable, is to condition the information flow instead of banning it directly. When dealing with explicitly diversity-concerned decision-makers, information could be released under the condition of a certain use, if active diversity measures are to be established. For example, in companies

---

64 Another limitation that can be noted from comparing both case studies is that, since the tool addresses the decision-making process only, it is unlikely to solve significant structural inequalities.
with a gender quota for the board of directors, but without one for interns, regulation could demand gender-blind resumes if gender discrimination was suspected. Then, if one of these companies wanted to go further in their diversity efforts and establish an affirmative action program for women in the workplace, applicants’ gender information could be allowed conditionally on the existence of such program.65

Anti-discriminatory privacy, however, would be effective at dealing with facially neutral screening rules set with ulterior motives. As the last part showed, the key to using privacy to fight discrimination is the identification and elimination of easily observable proxies (as defined in the previous section) for protected categories. Therefore, a facially neutral screening rule that covertly aims to discriminate can be addressed by identifying whether the seemingly facially neutral information used as the basis for a decision works as a proxy for a protected category. If it does, then the decision rule can be prevented by blocking the information flow, therefore achieving the same result than striking it down for its disparate impact, but all within a disparate treatment logic.

For example, in New York City Transit Authority v. Beazer, the Supreme Court decided that it was constitutional to not hire candidates under methadone treatment even if it created a disparate impact against black applicants; the decision hinged upon whether the claim was a Title VII one—and the dissent argued that it was.66 Similarly, in North Hudson Regional Fire & Rescue v. NAACP, 67 a fire department’s residency requirement was questioned for having a disparate impact against black candidates, but this time it was considered included under Title VII and the Third Circuit considered it to failed to have established business necessity.

An anti-discriminatory privacy rule preventing the information point on whether a candidate used methadone or

65 Of course, some degree of monitoring would be needed to ensure that the program was legitimate and not a covert way to find the information to install a discriminatory practice.
resided in the area would have avoided the efforts necessary to frame the case under Title VII, successful in one of these cases and unsuccessful in the other. In such a way, this tool allows us to address facially neutral forms of discrimination against minorities not protected by Title VII and discriminatory decisions outside of an employment relationship (and therefore not covered by its disparate impact branch), which are currently unprotected by the Equal Protection clause when interpreted under a disparate treatment logic.

This feature turns especially relevant to overcome the distinction between mutable and immutable traits. In our current legal context, courts starkly distinguish between personal traits based on whether they are mutable or immutable, which to a large extent traces to an artificial distinction between being a member of a protected category and behaving like one. Courts protect individuals from distinctions based on immutable traits such as skin color and chromosomes, but rarely so from those based on the behaviors typically associated with those groups. According to Yoshino, the underlying logic of this separation is that the behavior can be suppressed to blend into the mainstream.

In such context, more subtle kinds of discrimination exist towards groups that do not assimilate—either because they refuse to or cannot. Assimilation is the way to avoid this subtler form of discrimination and at the same time it is precisely its effect. “After all, the logic goes, if a bigot cannot discriminate between two individuals, he cannot discriminate against one of them.”

Privacy rules are an unexpected way out of the dilemma. What Yoshino should have said is that if a bigot cannot distinguish between two individuals, he cannot discriminate against one of them. When applied to concrete

---

69 Id.
70 Id.
71 Id.
72 Id.
discriminatory decisions, privacy rules can provide covering without covering; assimilation without assimilation.

B. Bias-based versus statistical discrimination distinction

The information dynamic identified above leads to consider that, despite bias-based discrimination and statistical discrimination are often considered to work very differently, and in many contexts they do, for the purposes of privacy rules they operate under the same dynamic. The same privacy-based preventive solution, therefore, can be applied to both.

In the “ban the box” example, for instance, whether employers were correct in believing that black men are more likely to have a criminal record (making it statistical discrimination) or they were mistaken (making it bias-based discrimination) is irrelevant for the purposes of eliminating the information that enables a discriminatory act. In fact, employers were correct on the general effect, but exaggerated its magnitude; the actual correlation was neither nonexistent nor as high as employers estimated—employers overreacted to an existing gap.75

This example illustrates how people often make assumptions and generalizations subconsciously, and these assumptions and generalizations about groups permeate their decisions about individuals both consciously and subconsciously. Moreover, people often make decisions based on facts that are neither entirely correct nor mistaken, but hold an exaggerated “kernel of truth.”77 Due to representativeness heuristics and confirmation biases, people also oftentimes do

---

74 See, e.g., Strahilevitz, supra note 21 at 373–375 (arguing as a pragmatic concern that disclosing previously private information will help antidiscriminatory efforts only as long as statistical discriminators significantly outnumber animus-based discriminators).
75 AGAN AND STARR, supra note 39 at 26.
not update priors as cleanly as neoclassical economic models might suggest. For these reasons, while a stark bias-based versus statistical discrimination distinction is useful for social sciences, and particularly economics, it is rarely a useful distinction for law, and it is not useful to determine which information flows should be blocked.

Privacy rules affect bias-based discrimination, however, with an added wrinkle as compared to statistical discrimination. For bias-based discrimination, privacy rules can more often than not effectively shelter the disfavored groups that they attempt to protect, but at the cost of not reducing those biases against them that generate the need for protective rules in the first place. This turns into a similar tradeoff as that of “covering” in which minorities’ civil rights face the tension of their short-term interests to blend in and their long-term interests to promote identity politics. The effectiveness will depend on the timing of the protection. For a minority with the ability to “cover” such as the LGBTQ community or religious minorities, an effective privacy rule that hides their status and protects them from a concrete discriminatory decision (being hired, or signing a contract such as having a cake baked for a wedding) will be easy to implement. However, one that does so in a long-term relationship will be more difficult for the individual and potentially more detrimental of the long-term interests of the group.

C. **Scope of the proposal**

While most examples used in this paper were restricted to employment discrimination, the anti-discriminatory privacy tool proposed here can be applied to a wider range of discriminatory practices. Employment discrimination is particularly relevant because employment is a clear way to distribute wealth and opportunities in our society. It also encompasses a normatively distinct set of decisions since it is covered by Title VII and therefore by disparate impact.

---

78 See Yoshino, supra note 65.
79 See *Id.* (explaining how different minorities cover differently).
Anti-discriminatory Privacy

2017

but employment is not the only socially relevant and institutionalized dimension in which discrimination takes place. These decision-makers could also be healthcare providers, law enforcement authorities, or university admission officers, among others.

The tool proposed can be used in any instance in which there is a decision-maker who has a certain degree of discretion and, among the data points available to decide, the law wants to prevent him or her from using one of them—such as minority status. People’s lives can be significantly affected by discrimination in some of these decisions, such as those in the insurance market. For example, there has been research showing that doctors are more likely to prescribe painkillers to white patients than they are to black patients—racial disparities in pain assessment lead to different treatment recommendations. While it is difficult to regulate this behavior through antidiscrimination law, one could imagine different ways to hide patients’ ethnicity in initial stages.

An increasingly relevant case of decision-makers that use data points deemed controversial is that of algorithmic discrimination. I apply the anti-discriminatory privacy framework to this context in order to illustrate how our understanding of and interventions to algorithmic discrimination could change once we view the interrelation between discrimination and privacy.

V. AN APPLICATION: ALGORITHMIC DISCRIMINATION

A. Unbiased decision-makers

The information dynamic described above finds itself further complicated when algorithmic decision-making comes

---

80 See Strahilevitz, supra note 25 (explaining that the New York housing market had traditionally discriminated musicians, because they are loud, and lawyers, because they know their rights too well).
81 Kelly M. Hoffman et al., Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites, 113 PROC. NATL. ACAD. SCI. 4296–4301 (2016).
into play. Algorithmic decision-making can have multiple benefits for our society. Algorithms are faster, and often more accurate, than human decision-makers. For many tasks, algorithms surpass our human abilities, and the set of those tasks is constantly expanding.

Most significantly, algorithmic decision-making implies that the importance of biases for discrimination seemingly decreases. While we humans are flawed and, even when well-meaning, host a wide array of implicit biases, algorithms are often presented as a fairer and unbiased decision-making agent.

The central issue with algorithmic decision-making is its opacity problem. A spam filter, for example, uses classifiers and predictors to determine whether an email is likely spam, but it cannot explain why it is such. Credit card fraud detection algorithms follow the same dynamic, as do credit scoring and

---


83 There seems to be a limitation to this principle: those processes that are automated within ourselves. If you ask a computer and a human to make a simple algebra calculation such as the square root of 87, the computer will obviously be able to do this faster. But if you grab and throw a ball and ask a human and a robot designed for this purpose to catch it, you will find that no robot exists so far that can do this as well as an average person. The distinction between one process and the other is normally referred to as a distinction between System 1 and System 2. See Keith Stanovich & Richard West, *Individual Differences in Reasoning: Implications for the Rationality Debate*, 23 Behav. Brain Sci. 645 (2000); Jonathan St B. T. Evans, *Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition*, 59 Annu. Rev. Psychol. 255–278 (2008). See also IN TWO MINDS: DUAL PROCESSES AND BEYOND, (Jonathan Evans & Keith Frankish eds., 2009).

84 There are lots of academic references that illustrate this position but perhaps the most interesting illustration is one of popular culture. In the movie *Moneyball*, Billy Beane (Brad Pitt) was hired to help choose baseball players for a team. At the beginning of the movie, there is a scene in which he observes the previous coaches choose the next player based on factors such as whether they look confident, are good looking, or have a girlfriend. Beane condescendingly suggests, instead, to have an “automated” process based on the performance statistics of each player. This is presented in the movie, and outside of it, as a fair and impartial process—and more often than not it is one.
loan decisions algorithms. In addition, the most accurate methods of machine learning seem to be the least explainable ones. Still, it should be noted that opacity, or inscrutability, is not inherent to an automated decision-making process compared to a non-automated one. Many human decisions seem inscrutable and opaque. Opacity and inscrutability are relevant because they are more frequent in algorithms.

Most importantly for our purposes, in the times of big data, we sometimes find a decision-making process with a discriminatory outcome where the decision-maker is not a person, but an algorithm that disfavors certain groups—even without intent. Increasingly, algorithms make socially consequential decisions. This makes it relevant to determine whether their decisions are discriminatory and, if so, how to prevent this outcome.

Opacity has been classified as either (i) intentional, as in trade secrets, (ii) a result of technical illiteracy, or (iii) a result of the characteristics and scale of a machine learning algorithm—a certain degree of opacity is inherent to the machine learning process. The three kinds of obfuscation require different kinds of responses if the results of the algorithm are discriminatory or generally deemed inadequate. For the first, disclosure of the decision-making process can be mandated, especially to determine whether antidiscrimination

86 [Cite Barocas’ A&E paper]
88 For example, COMPAS, described in the next paragraph, does not ask for race as an input. However, some of its data points, which have different weights, correlate to race. It is difficult to know for which ones this is the case because the algorithm is a trade secret.
89 Burrell, supra note 82.
law is being disregarded. For the second, expert auditing or general education to understand the process can be put in place. The latter is the most difficult to address; and for this one, anti-discriminatory privacy would be most relevant.

Northpointe’s risk assessment algorithm COMPAS, for example, is widely used to predict the likelihood that people who have been arrested will commit future crimes. It has recently been accused of producing racially biased results, having almost twice as many false positives for black defendants than for white defendants, and more frequent false negatives for white defendants than for black defendants. In addition, it was accused of being as accurate as a simple predictor based on prior count both on false positives and on false negatives—even when COMPAS and the prior count predictor disagree on 31% of cases.

A different risk assessment algorithm, sometimes used at the federal level to make probation decisions, was also found to give a higher average score of post conviction risk assessment to black people, creating disparate impact—even if


91 See Andrea L. Roth, Machine Testimony, 126 YALE LAW J. 1 (2017) (questioning the role of algorithms in trials and arguing that, in evidence law, more than accuracy, our worry should be whether the jury has enough tools to interpret the algorithm).

92 Julia Angwin Surya Mattu, Jeff Larson, Lauren Kirchner, MACHINE BIAS: THERE’S SOFTWARE USED ACROSS THE COUNTRY TO PREDICT FUTURE CRIMINALS. AND IT’S BIASED AGAINST BLACKS. PROPUBLICA (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last visited May 1, 2017); Jeff Larson Julia Angwin, Lauren Kirchner, Surya Mattu, HOW WE ANALYZED THE COMPAS RECIDIVISM ALGORITHM PROPUBLICA (2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (last visited May 1, 2017) (finding the false positives to be 23.5% for white defendants and 44.9% for black defendants, and finding the false negatives to be 47.7% for white defendants, and 28% for black defendants).

93 [Cite Chouldechova’s Algorithms and Explanations paper.]
the study concluded that bias was unlikely since 66% of the racial difference was attributable to criminal history.\textsuperscript{94}

Whether this conduct is still discrimination, for many, remains disputed. And the question can be extended to a multitude of dimensions. Is it desirable for an algorithm to call for people of some religious and ethnic minorities to be searched more often in airports? Is it acceptable if an algorithm impairs employment opportunities for overweight people? If it is not, who is responsible? People’s intuitions on these questions vary widely.

\textbf{B. How algorithms discriminate}

Discrimination in algorithms could be largely classified along two categories. There can be bias in the data that is used by the algorithm, and there can be bias among the people who create the algorithm that gets translated into the data-processing mechanism. The first is a bias in the input and the second a bias in the process.

The first kind, bias in the data, has been explored by prior research on this issue.\textsuperscript{95} If an algorithm is fishing in a section of the dataset that, for some reason, is not representative, it will produce a non-representative output. An algorithm can only be as good as the data that it is fed.\textsuperscript{96} This is a more complex version of a standard statistics problem that is solved by the law of large numbers when the sample is large enough. However, in algorithmic decision-making, it not always is, either because the algorithms uses an already existing biased database or works through a machine learning process in which it obtains biased examples. Oftentimes, the data fed to algorithms suffers of some sort of self-selection problem, like the one in the hypothetical example of predictive policing mentioned in section 2.A.

\textsuperscript{95} Barocas and Selbst, supra note 84.
\textsuperscript{96} See Solon Barocas & Andrew Selbst, Regulating Inescrutable Systems, (2017) (arguing that what is needed is not more data, but meaningful data).
Regarding bias in the process, it is important to note that even advanced algorithms are not entirely autonomous. Any machine learning algorithm that exists today works under supervised learning, where a person determines the output under a conditional probability ("given input X produce this output Y"). Under the examples mentioned before, supervised learning would consist on a person ordering “here is an email, output the probability of this email being spam” or “here is a potential tenant’s data, output the probability of this tenant defaulting on payment.” So, regarding intentionality, even if there is no human decision on the output, there is always a human decision on how the decision should be made.

Moreover, regarding data-gathering, there is a theory of human behavior in any algorithmic model—trained or untrained. People must select what features are important enough for the algorithm to consider in order to determine the output. Even advanced algorithms do not simply mine for random correlations.

Still, an algorithmic process could produce a result deemed undesirable, even when unbiased. An algorithmic process could produce a disparate impact on a minority due to correlations between pieces of information that were initially hard to predict. This traces a parallel between algorithmic

---

98 Because there are theories of human behavior behind every algorithm, disparate treatment is applicable. Cf. Kim, supra note 84 (arguing the opposite opinion: that disparate impact is not applicable to algorithms).
99 An algorithm that does this would be unsatisfying, as reality is plagued with spurious correlations. For example, the per capita consumption of mozzarella correlates with number of civil engineering doctorates awarded in the US (r=0.958), the per capita cheese consumption correlates with the number of people that died by becoming tangled in their bed sheets (r=0.947), the divorce rate in Maine is correlated with the per capita consumption of margarine (r=0.992). This leads to the inexorable policy conclusions of fostering mozzarella consumption but discouraging any other cheese, as well as margarine, but only in Maine. Tyler Vigen, 15 INSANE THINGS THAT CORRELATE WITH EACH OTHER SPURIOUS CORRELATIONS, http://tylervigen.com/spurious-correlations (last visited Apr 24, 2017).
discrimination and the human discrimination distinction between bias-based and statistical discrimination. Even if this distinction is unimportant for the purposes of regulating information flows, it illustrates how algorithmic discrimination and human discrimination equivalent for the purposes of anti-discriminatory privacy.

C. Privacy solutions to algorithmic discrimination

In section 3.A I mentioned that, as Strahilevitz explains, his conclusions on the relationship between privacy and discrimination as being in tension operate under a strict set of assumptions. I also argued in section 4.B. that those assumptions do not always apply to human decision-making. Interestingly, his assumptions on rational Bayesian updating do apply almost without exception to algorithmic decision-making, which makes his arguments more topical than ever before. Anti-discriminatory privacy rules are useful to address algorithmic discrimination not because they have, like humans do, a resistance to update priors, but because they have a facility to block individual data-points in the decision-making process. While it can be difficult to instruct a human decision-maker to disregard a visible fact, it is more feasible, even if not always simple, to code an algorithm to do so.

Algorithms can actually be productive for reducing discrimination. While it is true that human biases can be coded into the algorithm, the process of coding them makes them more explicit. This means that unconscious biases might be detected by the same programmer who holds them, or by subsequent reviewers, and not all biases will necessarily be transferred to the code. Moreover, that algorithms are coded makes it easier to regulate algorithmic decision-makers than human ones, absent trade secrets. While faulty logic is only figuratively codified in human decision-makers, it is literally codified in algorithmic ones. In some way, we demand more from algorithmic decision-making than we do for the human type, because we can.

---

100 See Strahilevitz, supra note 21; Strahilevitz, supra note 25.
Regarding information flows, algorithmic discrimination and human discrimination are not categorically different but have a parallel dynamic, as this part illustrated. Going back to the case studies, recall that, in the “ban the box” example, it was irrelevant for the purposes of eliminating the discriminatory act whether employers were correct or mistaken in believing that black men were more likely to have a criminal record. In the same way, whether a person made the ultimately discriminatory employment decision, or an algorithm did, is irrelevant for this purpose. This makes algorithmic and human discrimination equivalent with regards to anti-discriminatory privacy.

The privacy solution to algorithmic discrimination, in this sense, is to apply the same information policy to them than to any other kind of discrimination. That an algorithm can be biased too, and that it can produce a discriminatory outcome even without being so, as mentioned in the last section, traces back to the distinction between bias-based and statistical human discrimination from section 2.A. The two kinds of algorithmic biases that I argued can lead to discrimination (bias in the data that is used as an input and bias in the data-processing mechanism) trace back to the two conditions for anti-discriminatory privacy mentioned in section 3.A.

At a general level, therefore, applying anti-discriminatory privacy to algorithmic discrimination is important because it illuminates how privacy and antidiscrimination relate generally regarding information flows. Prior research by Barocas and Selbst and by Kim has helpfully examined how algorithms complicate our standard approaches to discrimination. With this proposal, I aimed to explore the opposite direction and show that algorithms also

---

101 By forcing us to confront these questions, algorithms impact our social norms regarding antidiscrimination by changing the context in which they operate and providing new counterexamples that challenge our existing intuitions. Algorithmic decision-making is telling of how we think about discrimination in general. This determines the extent to which our conclusions about privacy and discrimination under new technologies can be extended to reframe our general understanding of them.

102 See Barocas and Selbst, supra note 84; Kim, supra note 84.
illuminate the generic information dynamic of discriminatory decision-making, and they can affect how we think about discrimination also in the offline world. Algorithms have illustrated that biases are inevitable but they can be reduced when the environment forces them to be made explicit, that a decision process can only be as good as the data that is uses, and that there is always, at some level, a human decision-maker and a theory of human behavior.

VI. CONCLUSION

There are many ways to fight discrimination alongside Title VII claims. There is, for example, affirmative action in education, and establishing of quotas in executive boards. By exploring the information flows common in privacy norms and different kinds of antidiscrimination norms, in this paper I identify a new way to fight discrimination—by using privacy rules.

The main issue at hand when evaluating these antidiscrimination norms is to see whether there are proxies available for the information flow that the policy attempts to block. For an anti-discriminatory privacy rule to be effective, one must address those proxies as well.

This method is particularly relevant for algorithmic discrimination, a type of discrimination that continues to puzzle us. To some extent, the literature on algorithmic discrimination has so far provided legal solutions to an information problem. The tool that I propose here brings an information solution to the information problem that can work alongside the legal solutions previously proposed.

The contributions of this paper, in sum, are that it helps predict the effectiveness of antidiscrimination measures based on information restrictions and it explains how to design effective anti-discriminatory privacy rules.