Big Data and Compounding Injustice

Deborah Hellman

Draft, July 2020

I. Introduction

United States history is filled with injustice. One question for non-ideal theory, then, is how this prior injustice affects the obligations of people and institutions today. In very broad terms, there are several possibilities. First, each of us may have a duty to work toward the creation of more just institutions and the creation of a more just future.[1] Additionally, individuals and institutions may have duties to compensate victims of injustice for the harm they have suffered. And third, people and institutions today may be constrained in the way they interact with victims of injustice. Very roughly, these three ways of articulating how prior injustice may affect the rights and duties of people today divides the terrain into forward-looking obligations, backward-looking obligations and present-centered obligations. My focus in this article is on the third type of obligation – one which has attracted significantly less attention than the other two. I argue that the fact that a person has been a victim of prior injustice affects how others should treat her. In particular, this fact generates reasons that others should consider in deciding how they interact with her. This article's moral claim is that the fact that an action will compound a prior injustice counts as a reason against doing that action.

For ease of exposition, I call these reasons to act or refrain from acting so as not to compound prior injustice *The Anti-compounding Injustice principle* or ACI. This principle, if it exists, is likely be relevant to analyzing the moral issues raised by the increasing influence of so-called "big data" and its combination with the computational power of machine learning and artificial intelligence (AI). Decisions that rely on big data and machine learning are similar in kind to decisions which, also evidence-based, are grounded in less comprehensive information and where the processes used to analyze that data to make predictions about the future are less powerful. Where big data driven decisions differ is with regard to degree. If more types of decisions are data-driven in this way and these decisions are grounded in more data, then these new technological tools may compound more injustice than was possible before. If so, this is of moral concern.

A worry that is roughly along these lines likely animates much of the concern about the use of big data and algorithmic decision making in the context of consequential decisions like who is hired, offered a loan or released from prison. There are other worries too, of course, including those focused on accountability, transparency and concerns about fairness that are different from the worry about compounding prior injustice. My goal in this paper is to explain this particular worry and to defend the intuition that it is well founded.

[1] John Rawls, *A Theory of Justice* (Cambridge, MA: Belknap Press of Harvard University Press, 1971), 114-115.

II. Injustice and Big Data

There are different ways that prior injustices can affect the data that is used in decisions today. Broadly speaking, it is helpful to divide these ways into injustices that affect the accuracy of the data and those that do not. While not particularly elegant, one could call these *accuracy-affecting injustice* and *nonaccuracy-affecting injustice*. I describe each below. To date, data affected by *accuracy-affecting injustice* has attracted significantly more attention than data affected by *nonaccuracy-affecting injustice*. Yet, both are important and both types of injustice can be compounded when actors use data to make consequential decisions. In fact, as data collection practices are improved, one should expect that *accuracy-affecting injustices* will wane. It is thus especially important to remember that not all injustice that affects data affects the accuracy of the data and to interrogate how data infected by *nonaccuracy-affecting injustice* changes our moral obligations as well.

A. Accuracy-affecting injustice

Data can be flawed due to defective processes of data collection. Defective processes can result for ordinary mistakes, that is errors that are not morally culpable. Alternatively, defective processes can result from morally blameworthy data collection processes. Where data collection processes are defective due to injustice, this is *accuracy-affecting injustice*.

Consider first data that is flawed for reasons unrelated to injustice. So-called "measurement error" is ubiquitous and often unavoidable. Measurement error occurs because the trait of interest (T) is difficult to measure directly. Perhaps the trait is complex, like being a "good employee." In such a case, one must choose proxies for T that only approximate T. Measurement error can also occur because the trait is obscure, like commission of crime. In such a case, researchers choose proxies they can see and count, like arrests, that imperfectly track the trait that remain hidden. In addition, data are sometimes flawed because the collection process is skewed in some respect, inadvertently collecting information on only a subgroup of the class at issue. This list of ways that the data can inaccurately reflect the facts about people they purport to represent is just a sampling of potential problems. The point is that errors are routine and expected.

Errors that result from injustice are a subset of the errors that affect data. These too can result in a number of ways. Consider a few familiar examples. Suppose an employer would like to select employees and to do so uses an algorithm designed to predict job success. To build this algorithm, the employer uses data about features of past employees and information about their success on the job. However, the designer of the algorithm must determine what are the markers of "success." One possible measure on which the algorithm might rely is the performance reviews of managers. If these managers are biased in their assessments of prior employees[2] such that they rate women and minority employees who perform equally well as less good at their jobs

[2] This bias could result from animus or prejudice toward members of the protected groups or could result from implicit bias of which the managers are themselves unaware.

than they do white and male employees, then the error in the data will have a negative effect on these groups.

In such a case, the data is inaccurate. Women and minority group members really are as good as male and non-minority group members. The biased performance reviews affect the accuracy of the data. This is an accuracy-affecting injustice.

Policing provides another context in which injustice can affect the accuracy of data. Arrests for prior crime correlate with future criminal activity and so data on arrests may be of interest to police and policymakers. However, arrests are the product not only criminal activity but also of policing practices, which could be unjust.[3] If so, then using past arrests to predict to predict future crime relies on data that is flawed due to injustice.

When injustice affects data collection, as in these examples, it produces error. The women evaluated by sexist managers are better employees than the data suggests and reliance on this data thus inaccurately correlates sex with skill. African-Americans who are over-policed are not more likely to commit crimes than are individuals who are subject to less surveillance (or the differential is less extreme). Reliance on arrest statistics to predict future crime therefore overstates the correlation between race and crime.

Data collection practices can also be unjustly flawed due to skewed data collection processes. Consider, for example, the problem that facial recognition technology is better able to identify white faces than black faces.[4] Critics allege that this failing is due to the fact that the technology was trained on white people and that few African-Americans were included among the faces on which the algorithm was trained. If this failure to include sufficient minority members in the training data is due to injustice, this case provides another example of accuracy-affecting injustice.[5]

These examples of *accuracy-affecting injustice* are familiar and have attracted significant attention. When critics charge that data is biased, they usually have in mind these sorts of problems.[6] Were these biases cured or minimized, accuracy would improve. But errors would still remain. Some traits are difficult to measure, some facts are hidden and, of course, it is not possible to perfectly predict future events. However, if these injustices were cured, not only would accuracy improve but the distribution of the remaining errors would likely be different.

[3] *See e.g.* Sandra G. Mayson, "Dangerous Defendants," *Yale Law Journal* 127 no. 3 (Jan. 2018): 490-568 or Sandra G. Mayson, "Bias In, Bias Out," *Yale Law Journal* 128 no. 8 (June 2019): 2218-2300.

[4] *See* Joy Buolamwini & Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1-15

[5] In the first two examples, the *accuracy-affecting injustice* derives from behavior of managers and police and not from the researchers collecting the data themselves. In the third example, the *accuracy-affecting injustice* arises from the researchers own behavior. While this is a difference, it is not one that matters to the distinction I am drawing between data that is inaccurate as a result of injustice and data that is accurate and records prior injustice.

[6] U.S. Representative Alexandria Ocasio-Cortez, "Interview of Rep. Ocasio-Cortez at Blackout for Human Rights, MLK Now 2019," The Riverside Church in the City, Jan. 21, 2019, New York City, https://www.trcnyc.org/mlknow2019/ (interview with Rep. Ocasio-Cortez begins at approximately minute 16 and comments regarding algorithms begin at approximately minute 40); Julia Angwin et al.,"Machine Bias," *PROPUBLICA*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Rather than affecting previously disadvantaged groups disproportionately, the remaining errors would likely be spread more randomly (or perhaps evenly) across populations.

B. Nonaccuracy-affecting Injustice

A second way in which injustice can affect data occurs when the data accurately report facts about people where these traits themselves result from injustice. Injustice is likely to leave some people or groups with fewer skills, less wealth, poorer health and other traits that states, employers, lenders or others are interested in. For example, suppose that racial discrimination in employment is partly responsible for the fact African-Americans disproportionately occupy low wage jobs. Data, even accurate data, about earnings thus incorporates that injustice. Similarly, if political actors unjustly allocate funding between wealthy and poor neighborhoods and educational quality depends in part on funding, then some children will be less prepared for higher education or the workforce due to injustice. Data about educational attainment, though accurate, will contain or report that injustice.

In both the cases of *accuracy-affecting injustice* and *nonaccuracy-affecting injustice*, injustice infects the data. I use the word "infects" somewhat metaphorically as the way that injustice appears in the data is different in each case but nonetheless relevant. Injustice can cause the data to be inaccurate. Or injustice can cause the traits that the data report. These two ways in which injustice affects data are not themselves new. Before the advent of big data and machine learning, people collected data about others and made decisions on the basis of this data and in doing so relied on data that was infected by injustice in each of the ways I describe. What is new is the extent of these problems. The scope of big data and the power of machine learning augur a dramatic increase in the importance of data itself. For this reason, it is a moral problem that demands our attention. To date, scholars have focused on only one side of that problem: the compounding of *accuracy-infecting* injustice.[7] For that reason, in what follows, I focus especially on *nonaccuracy-infecting injustice* and the ways in which it too can be compounded.

III. Compounding Injustice

In this section, I first describe the phenomenon I term "compounding injustice" and then argue that the fact that an action will compound prior injustice counts as a moral reason that weighs against that action. At times, this reason will be overridden by other reasons, but the fact that an action will compound prior injustice should matter when individuals and institutions decide what policies to adopt.

It is important to distinguish this claim from others in the nearby vicinity.[8] Individuals have obligations to work toward the creation and maintenance of justice institutions in their

[7] Mayson, *Bias In, Bias Out*.

[8] Tommie Shelby identifies four principles that make-up non-ideal theory. These include: "(1) Principles of *reform and revolutions* are standards that should guide efforts to transform an unjust institutional arrangement into a more just one. (2) Principles of *rectification* should guide attempts to remedy or make amends for the injuries and losses victims have suffered as a result of ongoing or past injustice. (3) Principles of *crime control* should guide the policies a society relies on when attempting to minimize and deter individual noncompliance with what justice requires. (4) *Political ethics* are the principles and values that should guide individuals as they respond to social

society (including institutions that compensate victims of injustice) and to rectify those injustices for which they are responsible or from which they have benefited.[9] Assuming that a person has made reasonable efforts (however defined) to fulfill her obligations to do these things, the question I focus on is this: does the fact that she is interacting with a victim of prior injustice matter when she determines how she should act? My answer is that it does.

A. When does an action compound prior injustice?

When a person or institution (call her the "actor") interacts with a victim of injustice (call her the "victim") the actor sometimes *compounds* that injustice. What do I mean by that? In my view, if the actor engages with the prior injustice sufficiently, and also augments or entrenches that injustice, then the actor compounds the injustice and thus bears some responsibility for this worsening of the harm that the prior injustice now gives rise to This definition is rough. In particular, it is not clear what "engaging with the prior injustice sufficiently" means. In what follows, I will elaborate but this rough idea will get us started.

To crystalize the idea, consider the following example:

Life insurance and the battered women: Suppose Amy is a victim of domestic abuse. As a result, she is more likely to die during the upcoming year than a woman who is not an abuse victim. A life insurer, calibrating price for insurance policies to the likelihood that the insured will make a claim during the policy period, will therefore charge higher rates to Amy and other battered women than to similar others who are not abuse victims. Even if a battered woman leaves her abuser, she will be charged high rates by an insurer focused only on actuarial accuracy because women who leave are especially at risk of being killed.[10] An insurer who charges actuarially accurate rates to Amy will compound injustice.

In this case, the insurer charges Amy higher rates than the average person because of her status as victim of domestic abuse. In that sense, the insurer interacts or engages with the prior injustice. The insurer is not merely a bystander (though of course bystanders may have moral obligations to intervene in some contexts). Rather, the insurer takes an action that relies on, indeed uses, Amy's status as a victim in its rate-setting process. Second, for an actor to compound injustice the actor must augment or entrench the prior injustice. In Amy's case, if the insurer charges Amy higher rates than the average insured, the injustice that Amy suffered at the hands of her abuser takes on a second life. Not only has she suffered physically and emotionally, but that abuse will have financial repercussions as well.

injustices and that serve as the basis for criticizing the failure of individuals to promote just circumstances and to avoid complicity with injustice." Tommie Shelby, *Dark Ghettos: Injustice, Dissent, and Reform* (Cambridge, MA: The Belknap Press of Harvard University Press, 2016): 12. The *Anti-Compounding Injustice* Duty would fit within his "political ethics."

[9] *See e.g.* Erin I. Kelly, "The Historical Injustice Problem for Political Liberalism," *Ethics* 128 no. 1 (October 2017): 75-94 which argues that even ideal theory needs a principle of "historical redress" to address the racial bias embedded in society.

[10] *See e.g.* Deborah Hellman, "Is Actuarially Fair Insurance Pricing Actually Fair?: A Case Study in Insuring Battered Women," *Harvard Civil Rights-Civil Liberties Law Review* 32 no. 2 (Summer 1997): 355-412.

I use the word "compound" to describe the two morally relevant features because that term has two meanings that capture these two features. A "compound," used as a noun is a mixture. The harm that the insurer's pricing decision causes is the result of the mixture of the insurer's actions and those of her batterer. In addition, to "compound," used as a verb, is to intensify the negative aspects of something, just as the insurer does in augmenting the harm of the prior injustice.

In *Life Insurance and the Battered Women*, the actor used the fact that Amy was an abuse victim as the reason to charge her higher than average rates. If compounding injustice requires that an actor use the fact that another is a victim of injustice as the reason for her action, it will not apply all that often. Does compounding require the actor implicate herself in the injustice to this degree? To assess this question, consider another example:

Employment and unjust educational opportunity: Aliyah attended school in a poor neighborhood where the quality of education was weak. The schools were substandard due to an unjust distribution of resources. As a result, Aliyah lacks the skills sought by many employers. An employer who neglects to hire Aliyah due to her poor skills compounds the prior injustice.

In this case, the employer does not reject Aliyah because she suffered educational injustice. Rather it neglects to hire her because she lacks the relevant skills. Does this fact dilute the employer's responsibility for the harm of being rejected for employment sufficiently such that we should say that this harm, while part of the harm caused by the unjust educational opportunity, should not be attributed to the employer? I think this conclusion is too forgiving of the role the employer plays. While the employer is surely less responsible for this harm than is the state (to whom we might attribute the educational injustice), still the employer involves itself in the matter to a degree that it bears some responsibility for deepening the injustice. After all, the employer decides what criteria to use in making hiring decisions and how much weight to assign to each factor. These decisions are made against the backdrop of unjust educational opportunities. The employer cannot ignore the fact that some employees have poor skills due to injustice. The fact that its hiring policies will augment that injustice should count in the balance of reasons the employer considers when determining how to choose among applicants.

This example leads to a refinement of the definition of compounding injustice. When an actor takes the fact of prior injustice (as in *Life insurance and the battered woman*) or its immediate sequalae (as in *Employment and unjust educational opportunity*) as its reason for action, the actor involves itself sufficiently in the prior injustice to bear some responsibility for any extension of the harm of the injustice that its actions give rise to. In such a case, the actor compounds the prior injustice.

In these two examples, the nature of the prior injustice that is compounded is different. In Amy's case, the injustice is largely noncomparative. She has treated in a manner that conflicts with how she is entitled to be treated; she has been assaulted. In Aliyah's case, the nature of the injustice is largely comparative. She has been given less support for her education than have others, without adequate justification. Both noncomparative and comparative injustice can be compounded by subsequent actors. Moreover, sometimes particular injustices have both a comparative and a noncomparative dimension. While Amy has suffered an assault, the phenomenon of domestic violence is disproportionately suffered by women so there is also a comparative dimension to the injustice she has endured. And while Aliyah has been given fewer resources than other others, she may also have been given fewer resources than she is entitled to qua member of the political community (a noncomparative wrong).

The phenomenon of compounding injustice encompasses the compounding of both comparative wrongs, noncomparative wrongs and wrongs that have elements of each, as these examples make clear. In addition, it is a phenomenon that is common and does not depend on big data or fancy algorithmic tools. Yet, it is likely to occur when big data and machine learning are brought to bear, as the following example illustrates.

Lending and prior race discrimination: Darnell has been discriminated against in his search for employment. When employers see his application, his black-sounding name makes it difficult for him to be called in for an interview. If he does succeed in getting an interview, employers are less likely to hire him than they would be to hire a similarly qualified white applicant because his race affects how they perceive his credentials. As a result, Darnell has a lower salary than he would have if race discrimination had not affected his job opportunities. He is seeking a loan to buy a house. A lender is aware that its prior practice of using income as the central factor in determining whether to issue loans is likely to have a disproportionate effect on blacks and other racial minorities. As a result, the lender decides to deploy big data and machine learning to identify other factors that predict loan performance. Rather than using salary, the proposed new method uses credit history, health and marital status.

The substitution of these new factors may similarly compound the prior injustice of race discrimination in employment. Race discrimination led Darnell to have a lower salary than he would have had otherwise. This low salary may be related to his poor credit history. Alternatively, perhaps race discrimination in the context of lending itself has led to his poor credit history, as he may well have been a victim of predatory lending. In addition, race discrimination is likely to have contributed to his poor health, as evidence suggests that the stress of discrimination has tangible and measurable effects on health.[11] In addition, the discrimination in employment may well have has affected his marriage prospect as men with low salaries are less likely to be married than men with higher salaries.[12] Thus, race discrimination in employment and other contexts that Darnell experienced not only affect his salary. Rather it affects a myriad of other aspects of his life. While machine learning can offer the opportunity to find other traits that correlate with the target trait, there is a high likelihood that these other traits will also be caused by the injustice we hope to avoid compounding. When the lender uses these other traits, it also grounds its decision on direct effects of the prior race discrimination and augments the harm thereby caused. For this reason, the lender in this case also compounds injustice.

[11] David R. Williams, Jourdyn A. Lawrence, Brigette A. Davis, & Cecilia Vu, *Understanding How Discrimination Can Affect Health,* 54 HEALTH SERV. RSCH. 1374, 1383-84 (2019).

[12] Tara Watson & Sara McLanahan, *Marriage Meets the Joneses: Relative Income, Identity, and Marital Status*, 46 J. HUM. RES. 482, 482 (2011).

The factors that this hypothetical machine learning algorithm picks out to identify which borrowers will repay their loans was made up and simply used to make a point. More plausibly, lenders might substitute rent or utility payment history or perhaps more controversially education or occupation. It is easy to see how the race discrimination Darnell suffers (in employment and other contexts) might be related to his decision to attend a 2-year community college rather than a four-year institution or to work in particular occupations. Similarly, there is likely to be a connection between race discrimination and the ability to regularly pay rent and utilities. Just as *race* is not simply the pigment of one's skin, but instead denotes a cluster of social experiences, including a likelihood of being subject to discrimination and violence, so too race discrimination is likely to have direct effects on a myriad of experiences in a person's life. As a result, while it is surely possible that the use of big data and machine learning will be able to identify proxies for loan repayment that are not as tied to experiences of race discrimination as is salary or credit history, it is also entirely likely that this will not be easy to accomplish. In emphasizing the dangers that big data may simply compound injustice in much as the same way as conventional underwriting methods, I do not mean to suggest that lenders should abandon the goal of using these tools to expand access to credit to many people who have thus far been left out. Caution, however, is warranted.

B. What's wrong with compounding injustice?

The fact that the actor in each of the three scenarios described above involves him or herself with the injustice suffered by the victim and thereby augments or entrenches that injustice provides a moral reason for the actor to avoid that action. Why think so? In each of these cases, the actor is making the bad situation worse. It is her action that is worsening the harm of the prior injustice. While the actor is not responsible for the prior injustice, she decides how to act in the face of it. The fact that her action will deepen or entrench that harm matters morally.

Return to Amy and her life insurance purchase. In that case, the insurance company must decide *how* it responds to the prior injustice visited on Amy. While the insurance company is not responsible for what happened to Amy, its actions in setting prices will either augment and entrench this prior injustice or they will not. If it charges Amy higher than average rates because she is an abuse victim, she will suffer not only the original harm of abuse but also the harm of paying higher than average insurance rates. My intuition, which I hope the reader shares, is that the fact that Amy's risk status results from her being a victim of injustice should give the insurer pause with regard to its decision about how to set her rates. The insurer should pause because if the insurer charges her higher than average rates on the basis of the fact that she is a victim of domestic abuse, the insurer will add to and entrench the injustice she suffers.

To be sure, the insurer also has good reasons to charge the abuse victim higher rates than others. The company aims to charge prospective insureds actuarily accurate rates in order to maintain its business. What I am suggesting is that the insurer consider both the reasons in favor of charging actuarial accurate rates and the reasons against doing so in this case. The employer in Aliyah's case also has good reason to pass over her in hiring, as does the lender have good reason to decline to lend to Darnell. Indeed, in some cases the reasons that weigh in favor of the action will clearly outweigh the force of avoiding the compounding of injustice. For example, suppose that victims of domestic abuse are more likely to be abusive themselves.[13] If I am in a relationship with someone who is abusive and this person is also a victim of prior abuse, must a stay in a relationship with an abusive partner in order to avoid compounding injustice? In my view, the fact that breaking off the relationship will augment the harm already suffered by this victim is not a sufficiently strong reason to stay in the relationship, given the risk of harm to myself. Yet, at the same time, the fact that my partner's abusive behavior is likely traceable to the injustice he suffered is relevant. It counts as a reason to consider that weighs against breaking up or that counts in favor of giving my partner time to seek help for the trauma he suffered.

In this sense, my claim is relatively modest. I assert that the fact that an action will compound prior injustice should count as a moral reason against that action. Sometimes that reason will tip the balance against the action that will compound injustice but sometimes it will not. The ACI principle describes a reason with significant moral force, however. The fact that an action will compound prior injustice is not simply a harm that the action causes. Rather the actor should count the fact that through her action, the harm of prior injustice will be entrenched as a significant factor to weigh in considering how to act.

That said, in the case of *Life insurance and the battered woman*, I believe the ACI principle provides a sufficiently weighty reason such that an insurer should refrain from charging the abuse victim higher rates. One might worry, however, that if the insurer does charge Amy, and all abuse victims, average rates, despite the fact that they pose above average risks, then the insurer will lose money and may itself go bankrupt. That consequences seems too great a sacrifice to ask of the insurer. Perhaps the insurer could choose instead to spread these costs by charging all life insurance purchasers a bit more that it would otherwise. If only a single insurer makes such a decision, this insurer is likely to lose customers to other companies that make different choices. Thus, if one believes that all insurers ought to refrain from charging higher rates to abuse victims, it may be necessary to enact a law forbidding insurers from risk-rating on the basis of abuse victim status in order to avoid the fact that insurers who act properly will lose customers to those that do not and to avoid the competitive pressure against the morally correct action.[14]

So too in the case of *Lending and race discrimination*. Darnell is unmarried, has poor credit, and poor health because of unjust race discrimination. If the lender rejects his application for a loan on these bases, it compounds that injustice because the lender bases its decision on the direct effects of the prior injustice (compounding as mixing) and entrenches that injustice because not only is Darnell stuck in a low-paying job but is also denied a loan (compounding as making a bad situation worse).

This decision whether to "pay it forward," if you will, is a morally relevant act. When faced with the victims of injustice – both isolated and systemic injustice[15] – other actors must contend with the fact that their own actions may deepen and entrench that injustice. When an

[13] This example was suggested by an anonymous reviewer.

[14] Many states have adopted such laws. *See* Hellman, note 10.

[15] The prior injustice can be isolated or systemic and it can be non-comparative or comparative.

actor bases its decisions on either the victim's status as a victim or its direct effects and thereby makes the harm of the prior injustice worse, this compounding of injustice has moral significance and should count as a reason against doing the action in question.

The prospective employer and lender also have good reasons to decline to hire Aliyah and lend money to Darnell, just as the insurer also had good reasons to charge Amy an actuarially accurate rate. In order to assess whether the ACI principle provides a reason that is strong enough to outweigh these countervailing reasons, one must consider the sort of employment at issue and the type and likely effects of hiring a less qualified applicant. It may turn out that the ACI principle only makes a difference at the margins, meaning that it helps only those candidates whose skills and abilities are just slightly below their competitors (in the employment context) or slightly less that what would be required to get a loan (in the lending context), etc. If so, recognition of the force of the ACI principle will not be practically transformative. But it is important nonetheless for two reasons. First, the ACI norm provides a novel explanation for the common intuition that employers, lenders, admissions officials, etc. should favor slightly less qualified applicants from historically disadvantaged groups over others. Second, recognition of the importance of this reason, even where it is overridden, has moral significance.

The ACI principle is likely to be important in the era of bid data and machine learning as big data-driven decisions will often involve compounding of injustice. Employers, lenders, criminal justice officials and others use data and machine learning to develop algorithms to predict various traits of interest (employment success, loan repayment, recidivism). These target traits (T) are predicted using various proxies (P_1, P_2, P_3). If P correlates with T due to prior injustice, then using P to predict T may compound that injustice. Call this the direct case. Alternatively, if P correlates with a protected trait like race or sex due to prior injustice, then use of P to predict T may also compound injustice. Call this the indirect case. Consider an example of each mechanism.

Direct case: Suppose sex is correlated with credit worthiness and on this basis an algorithm used to determine loan eligibility used sex as one of the traits on which to predict credit worthiness (T). If sex is correlated of T because of prior sex-based injustice (women are paid less than men for doing comparable work, for example), then use of sex to predict T would compound that prior injustice.[16]

Indirect case: Suppose sex is correlated with income and income is correlated with credit worthiness. An algorithm uses income (P) to predict credit worthiness (T). If sex is correlated with P due to prior injustice, then use of P to predict T will also compound that injustice.

IV. Discrimination Law and Compounding Injustice

As the above discussion highlights, the anti-compounding injustice principle is well-suited to serve a justificatory role for discrimination law. Discrimination law forbids the use of

[16] Of course, use of sex in this manner is prohibited by law. I use the example to show how a reliance on sex in this context would compound injustice. This account thus provides a justification for the prohibition embodied in the law.

classifications like race and sex in most instances. There are many possible reasons for the law's disfavoring of race and sex-based classifications, including most prominently that the use of these classifications is often demeaning,[17] will generally disadvantage previously disadvantaged groups,[18] will constrain freedom[19] or fail to respect a person's autonomy[20] and will lead to bad consequences.[21] We should add to this list that the law's focus on particular protected traits is justified on the grounds that individuals and institutions have an obligation to avoid compounding injustice.

Race and sex will often correlate with other traits *because* prior injustice has effects in the world. Racial minorities and women are denied opportunities or channeled into certain roles. As a result, members of these groups have worse health, lower educational attainment, less wealth, etc. Were law to permit the use of race and sex as a proxy for these traits – so called rational or statistical discrimination – the law would permit injustice to be compounded. Instead, the law rejects the use of these protected traits as proxies for other traits, subjecting so-called "disparate treatment" on the basis of race and sex to a heightened burden of justification.[22] In doing so, the law forbids (nearly always) the compounding of that injustice.

The treatment of disparate impact discrimination – known outside the United States as "indirect discrimination" – is more equivocal. In U.S. law, disparate impact on the basis of race and sex is not recognized as a violation of the constitutional guarantee of equal protection[23] but is forbidden in certain contexts by statute.[24] In other jurisdictions, prohibitions on laws that produce a disparate impact on these basis, without adequate justification, are stronger and more comprehensive. A thoroughgoing commitment to the anti-compounding injustice principle would favor the latter approach.

The structure of discrimination law is also consistent with the ACI principle. According to the ACI principle, the fact that an action will compound prior injustice provides a reason that weighs against that action. However, this reason can be overridden by other considerations. For example, the fact that a prospective employee has poor skills due to prior sex or race-based injustice provides a reason to discount these poor skills when considering whom to hire.

[17] Deborah Hellman, *When is Discrimination Wrong?* (Cambridge, MA: Harvard University Press, 2008).

[18] Tarunabh Khaitan, *A Theory of Discrimination Law* (Oxford: Oxford University Press, 2015); Owen Fiss, "Groups and the Equal Protection Clause," *Philosophy & Public Affairs* 5 no. 2 (1976): 107–177.

[19] Sophia Moreau, "What is Discrimination?," *Philosophy & Public Affairs* 38 no. 2 (2010): 143-179. *See also* Khaitan, *Theory*.

[20] Benjamin Eidelson, *Discrimination and Disrespect* (Oxford: Oxford University Press, 2015).

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

[22] For example, in the U.S. disparate treatment on the basis of race and sex is subject to demanding judicial review under the equal protection clause of the Constitution and is treated in a similar fashion under statutory law. Civil Rights Act of 1964, Pub. L. 88-352, 78 Stat. 241 (1964).

[23] Washington v. Davis, 426 U.S. 229 (1976).

[24] Age Discrimination in Employment Act, 29 U.S.C. 623(a)(2) (2018) ("It shall be unlawful for an employer-to limit, segregate, or classify his employees in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's age."); Title VII of the Civil Rights Act of 1965, 42 U.S.C. § 2000c-2(k) (2018) (Lays out the burden of proof in disparate impact cases); and Fair Housing Act, 42 U.S.C. §§ 3604(a) & 3605(a) (2018).

However, there are also good reasons for an employer to prefer job applicants with good skills, both self-interested reasons and concerns for others. A legal regime that requires that policies which produce a disparate impact on the basis of race and sex be justified by "business necessity," for example, works to weed out policies that compound injustice without a sufficiently weighty reason to counter-balance this harm to the victims of prior injustice.

While legal regimes that treat direct and indirect discrimination similarly are perhaps most consistent with the ACI principle, U.S. law's more stringent treatment of disparate treatment on the basis of race and sex can be justified by a pluralist approach to the wrong of discrimination. If direct discrimination is more likely to have some other wrong-making feature, in addition to the fact that it will compound injustice, there is a reason to treat it more stringently. In my view, explicit race and sex-based classification is more likely to demean than is indirect discrimination. Thus, disparate treatment is especially likely to be morally unjustified because it usually demeans and often compounds injustice (without sufficient countervailing reason to do so) and therefore it should be prohibited in most instances. Disparate impact, by contrast, is far less likely to be demeaning, even though it often compounds injustice. For this reason, a legal regime that is more permissive of disparate impact makes some sense.

V. Complications

In this section, I consider two problems and complications with the account of the ACI principle. I first consider whether it is injustice in particular that one should avoid compounding. Perhaps one also has a reason to avoid compounding misfortune? Second, I consider the problem that the actor in question is unlikely to *know* that the correlation between the proxy trait [P] and the target trait [T] is caused by injustice or that the correlation between a protected trait (like sex or race) and the proxy trait [P] is caused by injustice. If so, how does this fact affect the account.

A. What about misfortune?

It is easy to craft a hypothetical like those I describe earlier that suggests that actors ought to avoid compounding misfortune. A person suffers a "misfortune" when she suffers a harm that is not traceable to a wrong done to her. By contrast, a person suffers an "injustice, when the harm she suffers results from a wrong. To explore whether actors ought to avoid compounding misfortune as well as injustice, consider the following scenario:

Risk assessment and family criminality: Suppose Brett grew up in a family in which crime was common. His father and older brother are both in prison. Suppose he commits a crime at age 18 for which he is convicted and incarcerated. In addition, suppose that Brett is later being considered for parole and data suggest that family criminality predicts recidivism. Should the fact that Brett grew up in this family count against him in a state's decision whether to grant him parole?

If your reaction to this case suggests to you that the state should resist using family criminality to predict recidivism,[25] then, in your view, we may also have a reason to avoid compounding misfortune. This may be correct. I limit my argument to the ACI principle because I am more confident that compounding injustice is a moral problem and it thus seems a good place to begin. However, I do not reject the view that avoiding the compounding of misfortune should also weigh in the balance of reasons.

B. Can we *know* that the correlation at issue is caused by injustice?

One might also worry about how one can know that the correlation at issue is caused by injustice. Take the *direct case* first. I argue that the ACI principle provides a reason for an actor to refrain from using sex as a proxy for a relevant target trait if sex is correlated with that target due to prior sex-based injustice. For example, suppose sex is correlated with credit worthiness. It is plausible to suppose that the reason that sex is correlated with credit worthiness is because women have lower salaries on average than men. It is also plausible to suppose that the reason that women have lower salaries is, at least in part, due to sex-discrimination in the workplace. Moreover, to the extent that women tend to choose lower paying fields than do men, this too could have its roots in gender-based injustice such that women are steered into fields that have lower pay and fields that attract large numbers of women are compensated less well than fields that attract more men.

But, one might worry, all this is supposition. Is supposition enough? The answer is yes. At the individual level, an actor deciding how to act must make a decision on the best evidence she has. If the best evidence suggests that she will compound prior injustice, then she has a reason to avoid that action. However, the strength of this reason will depend (in part) on how likely it is that her action will compound injustice. While at the level of public policy we may demand more certainty, the hypothesis that prior injustice has caused the observed correlations between protected traits like race and sex and various target traits is well-supported. The fact that sex-based classifications are disfavored in law is consistent with the view that correlations between sex and various target traits usually result from sex-based injustice.

Next consider the *indirect case*. Income is correlated with credit-worthiness. A lender thus uses income, in part, to determine to whom to lend. If sex is correlated with income because of sex-based injustice (for all the reasons described in the *direct case*), then use of income to predict credit-worthiness will also compound the sex-based injustice. Again, an actor does not *know* that sex is correlated with income due to prior sex-based injustice. However, she has good reason to suppose that that sex-based injustice is the cause. And a good reason to believe that injustice is the causes is sufficient to provide a reason to avoid such action.

There is a clear correlation between sex and income. What is contested is whether this correlation is caused by sex-based injustice. It might be tempting to rephase the claim that

[25] Jerrett Jones, Ellen Dinsmore, & Michael Massoglia, “Examining the Intergenerational Transmissions of Disadvantage: The Effects of Paternal
Incarceration among Young Adults in the United States,” *Paper Presented at Annual Meeting for the Population Association of America, Boston, MA, May 1-3* (2014), https://paa2014.princeton.edu/papers/142442Data on predictive tools that use it – COMPAS for example.

injustice causes the correlation between sex and income as a claim that female sex causes lower income. That is not my claim and, in addition, is one that is more controversial.[26]

VI. Implications for the use of big data and AI

In this section I consider whether use of big data and machine learning are especially likely to compound injustice or if instead these tools can help to avoid this problem. In addition, I explore the implications of the ACI principle for research in the field.

A. Is big data especially likely to compound prior injustice?

One might wonder what makes the ACI principle particularly relevant in the context of big data and AI. It is true that the moral problem of compounding injustice is not new. Indeed, as I argued above and have argued elsewhere,[27] discrimination law in the U.S. and elsewhere is justified at least in part by the fact that it stops the compounding of prior injustice. [28] Nonetheless, I worry that the scope of the problem of compounding injustice is likely to grow and the effect to deepen with the increasingly reliance on big data. This worry rests on an empirical claim and a speculative one at that. However, other scholars in the field have expressed similar worries.[29]

Big data driven decisions and actions are especially likely to compound injustice because the data collected is often the repository of prior injustice. The data records the effect of prior race and sex discrimination, of unjust distributions of resources and burdens. In addition, there are the problems of inaccurate data whose inaccuracy results from injustice. The heavy reliance on data infected with both accuracy-affecting injustice and nonaccuracy-affecting injustice will lead to significant compounding of prior injustice.

[26] *See*, e.g. Issa Kohler-Hausman and Lily Hu, “What’s Sex Got to do with Fair Machine Learning,” FAT*20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Jan. 2020): 513.

[27] Deborah Hellman, “Sex, Causation and Algorithms,” *Washington University Law* 98 (forthcoming, 2020) (arguing that sex-based equal protection doctrine can be explained and justified by the principle that states should avoid compounding prior sex-based injustice); Deborah Hellman, “Indirect Discrimination and the Duty to Avoid Compounding Injustice,” in *Foundations of Indirect Discrimination Law*, ed. Hugh Collins & Tarunabh Khaitan(Oxford: Hart Publishing, 2018), 105-121 (arguing that disparate impact liability (indirect discrimination) is best justified by a duty to avoid compounding prior injustice).

[28] Prior injustice that is unrelated to discrimination or mistreatment of legally protected groups is not addressed by discrimination law. However, specific laws do address the compounding of injustice. These include laws forbidding health and life insurers from risk rating on the basis of domestic abuse [**:***See* Emily C. Wilson, “Stop Re-Victimizing the Victims: A Call for Stronger State Law Prohibiting Insurance Discrimination Against Victims of Domestic Violence,” *Journal of Gender, Social Policy & the Law* 23 no. 3 (2015): 413-433 (“Forty-two states have passed laws prohibiting at least some kinds of insurance discrimination against domestic violence victims…”)], laws that prevent credit scores from being used when low scores are likely caused by the poverty of the scored individual. Mikella Hurley & Julius Adebayo, “Credit Scoring in the Era of Big Data,” *Yale Journal of Law & Technology* 18 no. 1 (2017): 148-216)

[29] *See e.g.* Borocas & Selbst, “Big Data”.

Perhaps, however, big data can be used to mitigate the compounding of injustice.[30] For example, a lender of the past might have relied on only a few traits to predict loan performance (say, income and credit score), each of which may have been heavily influenced by prior injustice. If big data and machine learning allow the lender to identify other traits that can predict loan performance equally well (or better) and are less likely to have been affected by prior injustice, perhaps these new tools will mitigate rather than exacerbate the compounding of injustice. This result is surely possible and important. If big data and machine learning can help to lessen the compounding of injustice, we should welcome this development. Indeed, research in the field should be aimed at exploring whether and how these new tools might be used to avoid or disrupt the compounding of injustice. That said, I worry that the opposite might occur, and want to encourage researchers and policy makers to be attentive to this risk.

However, as the discussion of *Lending and race discrimination* illustrates, in order for big data and machine learning to break the connection between past discrimination and future action, the algorithm must be explicitly directed to achieve this result. Researchers should look for traits that can predict the target trait while simultaneously lessening the disparate impact on historically disadvantaged groups. If machine learning can find new predictive tools that will not compound prior injustice, specifying this constraint should not significantly affect performance. If it does, this fact illustrates that the use of big data does risk significant compounding of prior injustice.

B. Implications for research

One implication for research, then, is to be attentive to the ways that new tools may exacerbate the compounding of injustice and to focus especially on looking for ways that new approaches might mitigate this harm instead. In addition, attentiveness to the moral problem of compounding injustice suggests that some approaches that equate fairness with accuracy will be inadequate.

Let’s return to the two types of injustice that might be compounded: accuracy-affecting injustice and non-accuracy affecting injustice. Accuracy-affecting injustice is marked by error. There is a gap between the proxy and the target (measurement error) and that gap is greater for one group than for another. Because African-Americans are policed more aggressively than whites, for example, arrests are a less good proxy for actual crime for blacks than for whites. Use of arrest data thus compounds the injustice of these policing practices. An approach that

[30] *See e.g.* Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig and Sendhil Mullainathan, ‘Human Decisions and Machine Predictions’, *Quarterly Journal of Economics* 133 (2018), pp. 237-93, (comparing human decision-making to machine decision-making in the context of bail and concluding that machines improved decisions while also reducing racial disparity in outcomes).

attempts to minimize the measurement error would avoid compounding this injustice.[31] One might aim then to increase accuracy or, as some argue, to treat truly like cases alike.[32]

But the injustice that we risk compounding via algorithmic tools reliant on big data is not only accuracy-affecting injustice. We also risk compounding non-accuracy affecting injustice. In such cases, the person ranked or scored or assessed by the algorithm really does lack the relevant skills, earn less money, is more likely to recidivate, etc. The moral problem lies in the fact that these differentials are due to injustice. A more accurate algorithm will not avoid this compounding injustice problem. Battered women *are* more likely to be killed during the term of the life insurance policy. In order to address the compounding of nonacccuracy-affecting injustice, the first step is to recognize this moral problem as distinct and one not addressed by efforts to achieve what some computer scientists call "individual fairness."[33]

For this reason, efforts by computer scientists to improve the accuracy of algorithms – while important for a variety of reasons – cannot fully address the moral issues they raise. The reliance on big data will continue to compound injustice, even if all the data on which they rely are accurate. Moreover, to the extent that increased computational power combined with massive amounts of data create novel opportunities to use these techniques in more ways and more pervasively, the degree to which these tools compound non-accuracy affecting injustice could significantly increase.

The research implications of this insight then are that computer scientists should not focus exclusively on increasing accuracy in their efforts to increase fairness. Treating like cases alike will not avoid compounding injustice when the injustice is unrelated to the accuracy of the data. Thus, efforts to clean up the data that is infected by accuracy-affecting injustice must be paired with efforts to dampen or ameliorate the ways in which reliance on accurate data can also compound injustice.

Conclusion

The use of big data together with machine learning may well allow the injustices of the past to affect the future to a much greater extent than ever before. It is therefore imperative to examine whether current actors have reasons to avoid compounding prior injustice. The injustice that could be compounded include accuracy-affecting injustices and the nonaccuracy-affecting injustices. I have argued for a reason to avoid compounding both types of injustice and have explained how this reason undergirds, at least in part, norms against discrimination. The upshots for big data and AI are twofold. First, while problems with biased data have attracted the most attention, this is only part of the problem that might concern us. Even where data are accurate, we risk compounding injustice. Second, efforts to improve the accuracy of data will only partially correct the problem at issue. When a protected trait is correlated with a legitimate target trait like having or lacking relevant job skills, we should go on to ask whether prior injustice

[31] Mayson, *Dangerous Defendants*

[32] *See e.g.* Cynthia Dwork et al., "Fairness Through Awareness," in *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (2012), 214, http://doi.acm.org/10.1145/2090236.2090255.

[33] Christopher Jung et al., Eliciting and Enforcing Subjective Individual Fairness,' ArXiv, May 25, 2019, https://arxiv.org/abs/1905.10660.

caused that correlation. In cases where prior injustice is likely the cause, the ACI principle provides a reason that weighs against using the target trait in the selection process.