# Fairness Under the Microscope of Big Data

Michael LoPiccolo*

December 7, 2016

*For ECS 3361.HN1, taught by Dr. Douglas Dow.

The age of big data is upon us: improvements in information technology, statistical methods, and machine learning promise to revolutionize decision making in every facet of our society. As software engineers tasked to build these decision making systems, we are best equipped to understand the danger and possibility inherent in these systems, and we must consider the social impact of our work. We must consider that impact in light of the long American history of discrimination, particularly along racial lines, and make special effort not to perpetuate it. We must engage with danah boyd's charge: "If you're building a data-driven system and you're not actively seeking to combat prejudice, you're building a discriminatory system."[1]

In this paper, we will focus on one particular danger: unintended racial proxy discrimination by classification systems which make economic decisions, such as credit, home loan, or employment decisions. Classification systems can take a data point holding information about a person and return a prediction of their suitability to the task (e.g. creditworthiness or employability). Big data systems use huge amounts of data about any one person to make predictions based on subtle patterns found from trawling huge datasets. How might such a system become racially biased? There are a few obvious ways, with obvious solutions. A programmer might have malicious, discriminatory intent. Such a programmer can be caught and fired. The data input into the system might be faulty or incomplete, misrepresenting minorities.[2] Such faulty data can be corrected. The system might be trained on race data and discover "surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality,"[3] using race as a helpful clue to someone's economic status. This is, of course, unacceptable, which is why legislation disallows considering race when making economic decisions.

However, we are concerned with a less obvious, more insidious way in which racial discrimination can arise, even with no bad intent, no faulty data, and no explicit information about race. It is *proxy discrimination*: even if race information is erased or ignored, it can be reconstructed by drawing connections between other pieces of data (location, web history, etc.) which serve as proxies. A machine learning system could easily learn to associate these proxies with lower economic status, and then assume that anyone exhibiting these proxies is, for example, more likely to

default on credit.[4]

There is no easy way to prevent this from happening. It is impossible to remove all proxies for race from the dataset; almost nothing will be left, as race influences so many different parts of life in our society.[5] And it's not just a bug in the system that we can fix: note that this inference, while unfair and discriminatory, will improve prediction accuracy, and is "rational." It's a result of big data systems being good at what they are intended to do – predict and discriminate. It is difficult to detect, because the models created by prediction systems are not comprehensible to humans. And it skirts current antidiscrimination legislation, because it never technically considers race. This is an extremely thorny issue; much has been written about it, but most work is still exploratory, nowhere near finding a complete solution. My goal will be to cast light on possible solutions by analyzing different perspectives: illuminating what a solution might look like, and what it certainly won't.

An important and immediate conclusion we can draw is that business-as-usual is untenable: old conceptions of privacy and old methods of protecting privacy according to those conceptions are insufficient to prevent the new dangers of big data. Historical conceptions of privacy are generally couched in terms of limiting access to personal data or controlling its propagation.[6] This suffices to prevent discrimination under the assumption that "assemble innocuous bits of information and you will have an innocuous assemblage; databases of non-sensitive information are non-sensitive databases."[7] Under this assumption, to prevent being discriminated against based on one's race, one merely needs to limit who knows one's race. But, in the world of big data, a heap of innocuous data can serve as a proxy for race. And trying to limit access to innocuous information is nearly impossible. Someone might be perfectly happy with my financial institutions knowing their mailing address to send statements, but still not want their location to be used as a clue to their race – the context of information use matters.

To properly diagnose the danger of big data discrimination, we need a more robust privacy framework. We find it in Helen Nissenbaum's framework of contextual integrity, which finds that privacy is violated when information flows inappropriately (in violation of contextual norms). An information flow is characterized by the actors (sender, recipient, and subject), type of information,

and transmission principle.[8] In contrast, previous access- or control-based privacy frameworks lack this context-sensitivity and specificity. They can only evaluate privacy violations at the point of access, not analysis or use. Additionally, the framework of contextual integrity allows us to consider the value of privacy to society, not merely to an individual.

The rigid access- and control-based frameworks fail to comprehend the new danger of proxy discrimination, so the historical methods that have been employed to protect privacy interests must also fail. Under old conceptions, data subjects' informed consent (such as by accepting a privacy policy) is sufficient to protect against privacy harms: the data subjects have voluntarily ceded access. Nissenbaum highlights two fundamental issues which cripple informed consent's ability to protect privacy against big data methods. First, for consent to be truly informed, the data subject must know not only what data they are giving access to, but what information can be gleaned from it. However, the value of big data techniques lies in their ability to discover previously unknown patterns and information. How can consent truly be informed when users do not know what information they are really giving away, hidden in supposedly innocous data?[9][10] Second, discrimination is a social harm, while consent is an individual decision. A "tyranny of the minority," of some consumers who consent to give up more data (rationally, as they benefit economically) can enable discrimination on the rest.[11] In fact, it could be rational for most members of society to consent to give up their data for economic benefits, even if discrimination meant some benefited much more than others.

We see that the access– and control-based conceptions of privacy are insufficient to meet the danger of big data discrimination, as they place privacy protection solely at the moment of collection and in the realm of the individual (consent-based methods). With the contextual integrity framework of privacy, we can analyze solutions which intervene at the points of data collection, analysis, and use ("implementation")..[12] These solutions are intended to prevent harm to society as a whole, not just to individual data owners. We will consider technical algorithmic remedies, as well as legal remedies, which attempt to forestall discrimination during the phases of analysis and implementation.

We begin by reviewing the literature on algorithmic fairness – searching for technical solutions to mathematically ensure fairness. As one might expect, there's no quick fix – but taking stock of what is possible and impossible will help us understand what trade-offs must be grappled with. Note that this field is quite new, and these algorithmic fixes are still quite far from the mainstream; it will take a sea change in opinion before they are integrated into the law, or even into common norms for companies.

A machine learning problem has this structure: a system is given some training data (e.g. students' data and how well they performed in college) and must find an optimal mapping (given an arbitrary student, predict as accurately as possible how well they will perform if admitted). We can add additional constraints, which may cause a slightly less optimal mapping to be found.

This is the approach taken by Dwork et al., who propose a mathematical definition of fairness. Their "fairness constraint" requires that "any two individuals who are similar with respect to a particular task should be classified similarly." If the best classification (maximal utility) does not meet this fairness constraint, a fair algorithm will need to settle for a classification with less utility but which does meet the fairness constraint. This approach is interesting in that it exhibits a "quantitative trade-off between fairness and utility"[13] – even though enforcing the fairness metric is not as strong as an affirmative action quota, and still may result in unequal outcomes between racial groups (disparate impact). The immediate problem with this approach is that, to enforce the fairness constraint, we need a metric to determine which individuals are actually "similar" with respect to the problem domain.

A paper by Zemel et al. takes a slightly different approach. It attempts to formulate the problem of fairness as finding an intermediate representation of the data points that hides all information about protected class membership (from the intermediate representation, it would be much more difficult to guess whether a certain data point was from a protected class or not), while preserving as much use of the information as possible. The use of an intermediate representation is particularly interesting, as we could make any existing algorithm fair by first converting the data into this intermediate representation.[14]

Both of these methods of fairness, however, have serious limitations – fundamental limitations, according to a paper by Friedler et al.[15] Their paper proposes a framework under which all fairness algorithms must fall, and shows that any algorithm for fairness will run into certain assumptions and trade-offs. The proposed framework defines algorithmic fairness in terms of the connections between three spaces. The *observed space* contains the data we have observed about a person. The *decision space* is the space of possible classification results. In one example they provide, the task is college admissions, the observed space might contain students' GPAs and test scores, and the decision space would contain admission or rejection decisions.[16] The framework adds one more, invisible space: the *construct space*, which contains those attributes we really want to use to make the decision (in the college admissions example, the construct space would contain students' expected amount of success in college). The construct space and observed space may differ: for example, a student might have great potential but poor SAT scores because of culturally biased questions.

With this formal framework, Friedler et al. can now precisely define what makes an algorithm fair. They propose: an algorithm achieves *individual fairness* if it maps people who are similar in the construct space to similar outcomes in the decision space. It is clear under this definition that it is easy to ensure individual fairness if we can guess the construct space from the observed space. They call this the WYSIWIG (what-you-see-is-what-you-get) assumption. Of course, not everyone is going to have the exact same performance in college if they have the exact same SAT scores and rank. But the WYSIWIG assumption doesn't require an exact mapping from the observed to construct space; there is room for error. The problem arises if there's error which consistently affects only certain groups (such as in our college admissions example, where, among students with the same potential to succeed in college, students of some cultural groups will still have lower SAT scores). This is structural bias: a distortion between groups when moving from the observed space to the construct space, contradicting the WYSIWIG assumption. If structural bias exists, we cannot achieve individual fairness.[17]

However, we would still like our algorithms to achieve *non-discrimination*, defined as: on av-

erage, the mapping from a group's construct space to the decision space is not skewed compared to any other group. In other words, the members of a group will not, as a whole, receive decisions which underestimate or overestimate their ability. Non-discrimination can be understood as analogous to individual fairness: individual fairness mandates that similar individuals get similar results, non-discrimination mandates that similar groups get similar results (and non-similar groups get *appropriately* non-similar results: there is no skew). Note also that non-discrimination does not imply that all groups will look the same in the decision space. For that to occur, we need an assumption: that all groups are actually similar in the construct space. This is referred to as the "we're all equal" (WAE) assumption. If the WAE holds, we can ensure non-discrimination by equalizing outcomes. Without assuming the WAE, we cannot achieve non-discrimination.[18]

Why would the WAE assumption not always apply? Consider again the example of college admissions, but with different construct spaces. If a group is routinely disadvantaged due to structural discrimination, causing children of the group to have less educational resources, they will score lower in a construct space measuring "academic ability." So, if an admissions board wants to reward students who are the most prepared, the WAE assumption would not be accurate. However, if the board considers the construct space to be "long-term potential," the WAE would be accurate, as there is presumably no difference in potential between the groups. So, the WAE may or may not apply depending on what values animate our decision.

To sum up, Friedler's framework, finds a set of (mathematically proven) limitations: Fairness can only be achieved under the WYSIWIG assumption. If WYSIWIG does not hold, structural bias exists, individual fairness is impossible, and we can only achieve non-discrimination if WAE holds. We can now critically analyze the previous papers in context of these limitations.

The Dwork paper can be read as adopting the WYSIWIG assumption. It assumes the existence of a metric to measure the actual distance between individuals with regard to a certain task, i.e. we can actually know the construct space. It is a useful algorithmic contribution, but certainly not a magic bullet; there is no way to achieve fairness without relying on these assumptions. Dwork admits that we might need to "make up" the metric, which of course requires us to make more

assumptions.[19] Dwork's method may help us to be more fair once we have made good assumptions, but math cannot solve the problem of fairness and make those assumptions for us.

The Zemel paper tries to be that magic bullet, maximizing everything: fairness, non-discrimination, and utility. We might expect such an approach to fail: under Orlin's framework, fairness cannot be achieved if structural bias exists (and in our society, it exists almost everywhere). Indeed, the Zemel paper lacks a clear goal. The very mathematical structure of their procedure reveals the unresolved tension in their approach. Their model has three quantified goals: retaining information (for utility of classification), hiding information about protected class, and achieving statistical parity between protected classes. However, because these goals are in conflict, it cannot maximize all of them together. Thus, the algorithm maximizes the sum of the three goals, each multiplied by an arbitrary weight ("hyper-parameter").[20] I find it completely unsatisfactory that we should be satisfied with "some" fairness and "some" non-discrimination. How are we to weigh these against each other, and against utility?

Having analyzed the technical remedies, we find that, while they may help to reduce some harm, technical remedies are essentially incomplete: they are always trade-offs, never magic bullets. To provide resolution to the problem of proxy discrimination, a technical solution must exist within a broader context of societal values, which inform the trade-offs it must make. However, these values cannot only be founded in professional codes and norms. Just making vague statements against discrimination will not be enough. Because of the "quantitative trade-off between fairness and utility,"[21] profit incentives will ensure the slow erosion of fairness, as those most willing to overlook discrimination will be rewarded. The only solution is collective action – enshrining a common understanding in the system of law by creating new law or reinterpreting the old.

The goal of antidiscrimination law in the United States can be viewed through two lenses: anticlassification and antisubordination. The anticlassification view holds that the goal of civil rights legislation is to prevent differential race-based treatment of any form. On the other hand, the antisubordination view holds that the goal of civil rights legislation is to dismantle a historical pattern of hierarchy.[22] The two have long existed in an unclear balance, but big data will force us

to make this balance explicit, according to a paper by Barocas.[23]

Essentially, anticlassification is about preventing procedural harms, while antisubordination is about preventing substantive harms. The goals do often synergize. If a minority group is in a position of economic subordination, it may be expected that group members will also often be victims of racial discrimination in employment, lending, and other areas. Preventing the procedural harm of racial discrimination will reduce the impacts of subordination and do substantive good towards breaking the hierarchy. Of course, it flows both ways: helping a group to overcome subordination will make race less of a factor in society, reducing the incidence of racial discrimination.

However, the two goals are not always in harmony. Many criticial disagreements in antidsicrimination jurisprudence can be located in the conflict between these two nexuses. In a recent Supreme Court decision, *Ricci v. DeStefano*, the Court gave judgement in favor of a group of white firefighters, citing anticlassification rationale. The firefighters had taken a qualifying test for promotion, but the city decided not to use the test results because of concerns that mostly only white firefighters qualified, and none of the African-American firefighters qualified. This was a procedural harm done to the white firefighters to prevent a substantive harm to the African-American firefighters.

The Court does not consistently uphold one rationale over another. In another example, *Parents Involved in Community Schools v. Seattle School District No. 1*, Justice Kennedy struck down a specific procedure of directly racially balancing schools, but noted that, often, procedural change is acceptable to prevent substantive harms, e.g. carefully choosing zoning lines to balance demographics.[24]

In the analog world, this tension can exist without fracturing, thanks to the ambiguity surrounding procedure. The decision to choose a different procedure, inspired by the desire to prevent harm to a racial minority, can also be reframed as trying to find a policy with more utility. For example, in *Ricci*, the city justified its decision to disregard the test results by citing not only concerns of racial discrimination, but of efficacy. As the argument went, the fact that the test had such a racially split result raised red flags that perhaps the test itself was biased – testing book knowledge

(literacy) instead of actual on-the-job competence. Big data classification systems lack this wiggle room. Barocas expresses a concern that "Not all of the mechanisms...seem to be amenable to procedural remedies...only after-the-fact reweighting of results may be able to compensate for the discriminatory outcomes."[25] The process of big data classification is intended to be as accurate as possible, and not subject to the human failing of unconscious bias. There is no utility justification to change the procedure, because the discrimination is not irrational[26] – we must explicitly encode a trade-off which violates anticlassification for antisubordination.

We see that, far from providing a ready-made collective-action solution to our animating problem, the law may even hinder, or at least confuse, the attempts of individual engineers or companies to limit discrimination. If anticlassification triumphs over antisubordination as the dominant rationale for civil rights jurisprudence, a quite likely shift,[27] algorithmic remedies for substantive harm may be considered themselves a form of illegal classification.[28]

To sum up, we now have a clearer view of the twisting, foggy road to a solution for proxy discrminiation. Previous understandings of privacy nor civil rights law will be enough to protect minorities. Any technical solution will represent a trade-off between fairness and utility, and must be informed by a strong understanding of the values at stake. This understanding will be a difficult brokering and so will not be legitimate or effectual if it is only expressed in professional codes and norms. Rather, it must be enshrined in the law, requiring us to find a new understanding of the meaning and extent of civil rights in the age of big data, to settle the newly-quantified conflict between anticlassification and antisubordination. Only then can we find a proper technical solution to, as best we can, protect the vast, contradictory, beautiful complex of values which makes America truly great.

This will be a difficult, thorny path, but we should take heart. Big data is in one sense a wholly new problem, but in another sense it is merely revealing the contradictions that have always haunted antidiscrimination efforts. We should look to the opportunity to resolve these contradictions with both fear and hope. Big data carries with it the promise to reduce irrational discrimination and improve life for minorities;[29] shutting down all big data systems out of fear would cause even more

harm to minority groups. However, we should also not shut down this conversation. We should keep working to recognize and admit the problems and dangers of big data, and work towards resolutions, difficult as it may be.

# Notes

[1]danah boyd, "Be Careful What You Code For," *apophenia (blog)*, June 2016, http://www.zephoria.org/thoughts/archives/2016/06/14/be-careful-what-you-code-for.html.

[2]Solon Barocas and Andrew D Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016): 680-687.

[3]Ibid., 671.

[4]Ibid., 691-692.

[5]Ibid., 721.

[6]Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford, CA, USA: Stanford University Press, 2009), 69-71.

[7]Ibid., 203.

[8]Solon Barocas and Helen Nissenbaum, "Big Data's End Run around Anonymity and Consent," in *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, ed. Julia Lane et al. (Cambridge University Press, June 2014), 47.

[9]Ibid., 59-61.

[10]Also see Tal Z Zarsky, "Desperately seeking solutions: Using implementation-based solutions for the troubles of information privacy in the age of data mining and the internet society," *Me. L. Rev.* 56 (2004): 43-44.

[11]Barocas and Nissenbaum, "Big Data's End Run around Anonymity and Consent," 61-63.

[12]Zarsky, "Desperately seeking solutions: Using implementation-based solutions for the troubles of information privacy in the age of data mining and the internet society," 17-33.

[13]Cynthia Dwork et al., "Fairness Through Awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12 (Cambridge, Massachusetts: ACM, 2012), 2, doi:10.1145/2090236.2090255.

[14]Although, the algorithm will be less lossy if the intermediate representation is customized to each situation; if we make one intermediate representation for many different classifiers to use, more utility will be lost.

[15]Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016,

[16]Or, the decision space could contain some numerical value predicting a student's performance in college, and those with high enough scores would be admitted.

[17]Because, under structural bias, we cannot tell which individuals are similar in the construct space, in order to give them similar outcomes!

[18]If we simply give all groups equal outcomes when they are not equal in the construct space, we will overcompensate and discriminate against the group which was better in the construct space.

[19]For example, Dwork offers that adding points to the SAT scores of those in disadvantaged groups is one way to

"make up" a metric. We can understand this as trying to reduce some of the structural bias between the observed and construct spaces, to get closer to WYSIWIG. Dwork et al., "Fairness Through Awareness," 3.

[20]Richard S Zemel et al., "Learning Fair Representations.," *ICML (3)* 28 (2013): 3.

[21]Dwork et al., "Fairness Through Awareness," 2.

[22]Helen L Norton, "The Supreme Court's post-racial turn towards a zero-sum understanding of equality," *William & Mary Law Review* 52 (2010): 206-207.

[23]Barocas and Selbst, "Big Data's Disparate Impact," 675.

[24]Norton, "The Supreme Court's post-racial turn towards a zero-sum understanding of equality," 212-213, other?

[25]Barocas and Selbst, "Big Data's Disparate Impact," 715.

[26]Sometimes we can have a justification to change procedure – if we correct for bad input data, or change the social values informing our procedure, as in the example of deciding the construct space for a college admissions decision, we can also mitigate racially harmful results without sacrificing utility. However, we are not focused on the first case, and taking the second option still requires us to classify based on race.

[27]Norton, "The Supreme Court's post-racial turn towards a zero-sum understanding of equality."

[28]Barocas and Selbst, "Big Data's Disparate Impact," 725.

[29]Ibid., 676.

# Bibliography


Barocas, Solon, and Helen Nissenbaum. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and HelenEditors Nissenbaum, 44–75. Cambridge University Press, June 2014.

Barocas, Solon, and Andrew D Selbst. "Big Data's Disparate Impact." *California Law Review* 104 (2016).

boyd, danah. "Be Careful What You Code For." *apophenia (blog)*, June 2016. `http://www.zephoria.org/thoughts/archives/2016/06/14/be-careful-what-you-code-for.html`.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness Through Awareness." In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ITCS '12. Cambridge, Massachusetts: ACM, 2012. doi:`10.1145/2090236.2090255`.

Friedler, Sorelle A, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im) possibility of fairness." *arXiv preprint arXiv:1609.07236*, 2016.

Nissenbaum, Helen. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA, USA: Stanford University Press, 2009.

Norton, Helen L. "The Supreme Court's post-racial turn towards a zero-sum understanding of equality." *William & Mary Law Review* 52 (2010): 197.

Zarsky, Tal Z. "Desperately seeking solutions: Using implementation-based solutions for the troubles of information privacy in the age of data mining and the internet society." *Me. L. Rev.* 56 (2004): 13.

Zemel, Richard S, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. “Learning Fair Representations.” *ICML (3)* 28 (2013): 325–333.