Algorithmic Fairness – Algorithmic Discrimination

Marcello Di Bello - ASU - Fall 2021 - Week #10

Our goal today is to understand whether, and if so to what extent, the jurisprudence about discrimination can help us make sense of the notion of algorithmic fairness. The answer—according to Barocas and Selbst[1]—is "it is not very helpful", but it is nevertheless useful to understand why exactly.

The next question is, if current discrimination law is no good for identifying algorithmic discrimination, what would an adequate legal definition of discrimination look like, one that correctly captures algorithmic discrimination? Is this even possible?

To answer this question, we need to ask another, even more fundamental and difficult question. What is algorithmic discrimination?

We have been asking this question throughout the semester under different names, focusing on the positive (algorithmic fairness) rather than the negative side (algorithmic discrimination). These notions may come apart. Is algorithmic fairness a broader notion than lack of algorithmic discrimination?

Discrimination in data mining

Barocas and Selbst provide a taxonomy. Their taxonomy include things that some may not agree count as algorithmic discrimination (or discrimination in data mining, in their terminology). According to this taxonomy, data mining is discriminatory at several junctures:

1. The definition of the target variables (such as "creditworthy")
2. Training data: 2a. labeling and 2b. representativeness of the data
3. Feature selection for making predictions
4. Patterns in the data
5. Masking

First four items correspond to aspects of the data mining: 1. target variables are defined; 2. data are collected and labelled; 3. predictors are selected; 4. correlation are found in the data.

For each of the 5 items, discrimination is evidenced by an outcome that goes to the detriment of a protected group. Here is an illustration. Suppose data mining finds patterns of higher loan default rates that correlate with lower academic performance and lower socio-economic status. Socio-economic status and academic performance is then used to predict the target variable "creditworthiness". The algorithmic predictions have a disproportionate adverse effect on protected groups and exclude them from access to credit.

We have seen this vicious cycle a number of times. Deborah Hellman calls it compounding injustice.

Barocas and Selbst discuss the target variable "creditworthiness":

> creditworthiness is a function of the particular way the credit industry has constructed the credit issuing and repayment system. That is, an individual's ability to repay some minimum amount of an outstanding debt on a monthly basis is taken to be a nonarbitrary standard by which to determine in advance and all-at-once whether he is worthy of credit. (p. 679)

Do you agree with this constructivist point of view? For them, crediworthiness is not tracking anything real except the interests and expectations of lenders.

[1] Big Data Disparate Impact, *California Law Review*, 2016, 104:671-732

Title VII Employment Discrimination

Let's not turn from data mining to antidiscrimination law in employment. There are two main legal doctrines.

A. Disparate treatment

A successful claim of disparate treatment discrimination in employment is established, as follows:

STEP 1: plaintiff makes a *prima facie* case showing that a similarly situated person who does not belong to the protected group (race, gender, etc.) would have been treated differently.

STEP 2: the defendant (employer) responds with a legitimate, non discriminatory reason for the treatment.

STEP 3: plaintiff establishes by preponderance of the evidence that the reason given by the defendant is pre-textual.

The assumption here is that, if the reason is pre-textual, the emplyoyer's true reason for the decision must be the intention to discriminate.

B. Disparate impact

A successful claim of disparate impact discrimination in employment is established, as follows:

STEP 1: plaintiff makes a *prima facie* case showing that members of a protected group are disparately impacted (say they are hired for high paying positions at lower rates).

STEP 2: defendant (employer) responds that the employment practice has a job-related purpose ("business necessity defense"). The interpretation of this second step varies. It can be more or less stringent, depending on how "necessary" the employment practice in question should be for the business purpose.

STEP 3: plaintiff establishes that the employment could have achieved the above job-related purpose with an alternative employment practice that would have had no disparate impact.

Question: Are the forms of data mining discrimination seen earlier included in disparate treatment or disparate effect? Barocas and Selbst think that they are not. Why do they think so?

Here is a clarification about the last step and the second step:

> it is hard to imagine a business practice that is "necessary" while there exists a less discriminatory alternative that is just as effective. If business necessity or job-relatedness is a less stringent requirement, though, then the presence of the alternative employment practice requirement does at least give it some teeth. (p. 706)

Predictive Algorithms and Title VII

Can we use employment discrimination law—as described above—to show that algorithms discriminate? Barocas and Selbst think that predictive algorithms are likely to overcome discrimination challenges under Title VII.

Disparate treatment rests on intentional discrimination, but most algorithmic discrimination is unintentional. So this is a non-starter.

Question: Why is algorithmic discrimination unintentional? What about masking? What do Barocas and Selbst say about that?

Disparate impact is more promising for establishing algorithmic discrimination. To answer the charge of discrimination, the employer should show that the employment practice based on algorithmic recommendation serves a business purpose (STEP 2):

> Once a target variable is established as job related, the first question is whether the model is predictive of that trait. The nature of data mining suggests that this will be the case ... The second question asks whether the model adequately predicts what it is supposed to predict. (p. 707)

If the employer shows that the deployment of the algorithm has a business purpose, the plaintiff can still invoke the alternative employment practice prong (STEP 3).

> ... there is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense. Models trained on biased samples, mislabeled examples, and limited features, however, might trigger liability under the alternative employment practice prong. If a plaintiff can show that an alternative, less discriminatory practice that accomplishes the same goals exists. (p. 709)

Are reforms possible?

If current law about employment discrimination is unable to capture the distinctive ways in which algorithms discriminate, how should the law be reformed? This is not going to be easy.

Internal difficulties

Recall the five ways algorithms can discriminate. These forms of discrimination cannot be easily declared illegal as a matter of law.

Question: Should impossibility theorems be added to the internal difficulties?

1. The definition of the target variable can be a source of algorithmic discrimination. But target variables are necessarily open to interpretation, disagreement, subjective decisions. The law can hardly regulate this process.

2. Biased labels in the historical data are another sources of algorithmic discrimination. Such biases are widespread. Yet dispensing with historical data would end data mining itself.

3. Predictors that are not fine-grained enough are another source of algorithmic discrimination. The granularity of the predictors could be improved by collecting more data, but when is enough enough? We can apply a cost-benefit analysis model, assessing:

> ... (1) whether the errors seem avoidable because (2) gaining access to additional or more granular data would be trivial or (3) would not involve costs that (4) outweigh the benefits. This seems to suggest that the task of evaluating the legitimacy of feature selection can be reduced to a rather straightforward cost-benefit analysis. (p. 719)

The problem is, this cost-benefit model might recommend that disadvantaged groups suffer disproportionately, say, false positives decisions so long as the costs of balancing these errors across groups outweigh the benefits.

External difficulties

As some argued, two rationales underlie antidiscrimination law:[2]

- *Anticlassification*: the goal of antidiscrimination law is to ensure that protected classifications are not used as criteria, reasons, grounds for a decision. This is a form of procedural fairness, loosely connected with a prohibition against disparate *treatment*.
- *Antisubordination*: the goal of antidiscrimination law is to ensure that substantive inequalities and group disadvantages are eliminated. This is a form of substantive fairness, loosely connected with a prohibition against disparate *impact*.

These two rationales can be invoked to make sense of our intuitions about what we find discriminatory. Here is a challanging case:

> ...should the law permit a company to hire no women at all—or none that it correctly predicts will depart following the birth of a child—because it is the most rational choice according to their model? The answer seems obviously to be no. But why not? (p. 727)

Disparate treatment doctrine avoids discrimination if there is a non-discriminatory rationale for the decision. So disparate treatment is unlikely to answer the question in the right way. Could disparate impact (antisubordination) do better?

> ... the prohibition would have to rest on a substantive commitment to equal representation of women in the workplace. (p. 727)

Barocas and Selbst conjecture that endorsing an antisubordination reading of discrimination law will help in making sense of algorithmic discrimination. However, this will not be easy, especially because the US Sup Ct has disregarded the use of disparate impact (the closest notion to antisubordination) for constitutional questions about discrimination not related to employment practices. For example, an elaborate statistical analysis – showing that death penalty decisions in Georgia disproportionately targeted African Americans, controlling for several variables – was not enough to convince the Court that the system violated equal protection (see McClensky v. Kemp (1987)).

What we discussed was limited to title VII discrimination. But the other legal source for discrimination is the Equal Protection clause of the 14th Amendment to the US Constitution.

[2] On this distinction, see Fiss, Another Equality, *Issues in Legal Scholarship: The Origins and Fate of Antisubordination Theory*, 2004, Article 20