Algorithmic Fairness – Fairness v. Welfare

Marcello Di Bello - ASU - Fall 2021 - Week #11

Our goal today is to examine algorithmic fairness from the point of view of utilitarianism. We examine an article by Aziz Huq[1] who argues that efficiency and equity coincide, at least as far as algorithmic decisions are concerned. The unifying question for today discussion can be put as follows: what does utilitarianism and more generally consequentialism have to say about algorithmic fairness? As we will see, they have quite a bit to say!

[1] Huq (2019), Racial Equity in Algorithmic Criminal Justice, *Duke Law Journal*, 68: 1043-1134.

Equal protection jurisprudence

As a preliminary point (PART II), Huq shows that equal protection jurisprudence based on the 14th Amendment to the US Constitution is unhelpful for understanding what algorithmic fairness requires.. This jurisprudence focuses on two main ideas: (i) limiting the use of racial classification; and (ii) banning racialized intention. A third idea—(iii) limiting racial disparities or disparate impact —has been mostly aside in equal protection jurisprudence.

Huq's analysis of equal protection jurisprudence complements the discussion about Title VII antidiscrimination employment law from last week. See Barocas and Selbst, Big Data Disparate Impact, California Law Review, 2016, 104:671-732

(i) Racialized intent

Any differential treatment that is due to racial animus or discriminatory intent is clearly illegal. The use of race as a facto in a decision is not automatically prohibited, but must be justified with a clear rationale, an inquiry known as "strict scrutiny". The Supreme Court has allowed the use of race in specific contexts under strict scrutiny, for example, in college admissions for the purpose of fostering diversity.[2]

[2] See Fisher v. University of Texas at Austin, 570 U.S. 297 (2013) and Fisher v. University of Texas at Austin, 579 U.S. ___ (2016).

Algorithms hardly have a racially discriminatory intent. Perhaps, the argument could be that algorithms rely on racially biased historical data and intentionally using such algorithms counts as having a racially discriminatory intent. But biased data do not seem enough to trigger a constitutional violation of equal protection. Huq writes:

> "Even if flawed training data were identified, it seems unlikely that its tainted nature could suffice to establish a constitutional concern in practice. Any moderately competent municipality found using flawed data would ... defend its decision as the best option given historically shaped constraints. Because a constitutional violation cannot be shown unless the state relied on race as a ground of decision, as opposed to acting in spite of race, this defense would likely succeed." (p. 1093)

(ii) Racial classification

Racial classification is usually prohibited, partly on the ground that an explicit use of race as criterion in a decision sends publicly a demeaning message and entrenches racial stereotypes. But the use of race in algorithms—if race is used at all—is usually opaque. No demeaning message can be clearly identified.

Further, racial classification is not always prohibited. For example, race is routinely used as part of a suspect description in criminal investigations. Federal Appellate courts have not objected to this practice and the US Supreme Court has declined to intervene.[3]

[3] See, for example, Brown v. City of Oneonta 195 F.3d 111 (2d Cir. 1999), cert. denied, 534 US 816 (2001).

Huq thinks that the use of race as one of the predictors in an algorithmic classifier is akin to the use of race in a suspect description:

> "a classifier based on training data is akin to a suspect description of a familiar sort, insofar as both are predicated on historical facts about crime. ... an advocate of algorithmic criminal justice might note that human observers are more likely than a machine to err in their deployment of race as a signal of criminality than an algorithm." (p. 1096)

Do you agree with this parallelism between race as a predictor and race as feature in eyewitness description?

(iii) Racial disaparity

What about *disparate impact* as invoked in Title VII employment discrimination cases? Evidence of disparate impact against a protected group is enough to make a *prima facie* case of discrimination; see e.g. Hazelwood School District v. United States (1977). This applies to sectors such as employment and housing. The criminal justice system, however, seems exempt.[4]

[4] In McClensky v. Kemp (1987), the US Supreme Court ruled that disparate racial impact is not enough to establish a constitutional violation. 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.

Today and back then

Huq makes a historical comment:

> "Current doctrinal approaches to constitutional racial equality ... were configured in the context of judicial efforts to dismantle educational segregation in the Jim Crow South and then during a political backlash to the Civil Rights Movement ... the legal conception of racial discrimination as a matter of intention or classification would reflect judicial concern with the discretionary choices of the police officer, school board president, or state legislator—that is, the modal problems presented by mid-century civil rights law." (p. 1101)

Today's context is much different:

> "A set of tools developed for a regulatory world of dispersed state actors, occasionally motivated by naked animus, cannot be mechanically translated into a world of centralized, computational decision-making." (p. 1103)

Cost/benefit analysis

If the legal framework is either obsolete or inapplicable, what else is there? In PART III, Huq puts forward a cost/benefit framework:

> "the key question for racial equity is whether the costs that an algorithmically driven policy imposes upon a minority group outweigh the benefits accruing to that group." (p. 1111)

Here Huq is applying the maxim that we should select the course of action in which the expected costs outweigh the expected benefits.

The expected cost (or benefit) of an action is the sum of the costs (or benefits) of the possible outcomes associated with the action, where the cost (or benefit) of each outcome is weighed by the probability of the outcome. So the expected cost of <taking an umbrella> is the sum of the cost of taking an umbrella when it does not rain (outcome 1) and the costs of taking an umbrella when it does rain (outcome 2), each outcome weighed by its probability.

Spillover costs and two decision thresholds

A narrow view considers just immediate benefits (say, increased public safety) and immediate costs (say, unwarranted detention). But besides immediate costs, there are also broader externalities:

> "[they] take many forms, including the effect of high incarceration rates on black communities and children as well as the social signification of race as a marker of criminality." (p. 1113)

These spillover costs on family life, employment and racial stigma, are likely to disproportionately affect minorities. As Huq writes:

Why does this claim hold?

> "the spillover costs of coercion of minority individuals for the minority group will be greater on a per capita basis than the costs of coercing majority group members." (p. 1113)

This observation has an important consequence:

> "accounting for both the immediate and spillover costs of crime control when its immediate benefits are small conduces to a bifurcated risk threshold—one rule for the majority, and one for minority." (p. 1131)

In other words, if decision thresholds should be set at a socially efficient level balancing the costs and benefits of pre-trial coercion, and if the costs of coercion are higher for blacks than for whites because of uneven spillover costs, the standard for imposing coercion must be *more stringent* for blacks than for whites, other things equal.

The emphasis on spillover costs has the merit of broadening the discussion about algorithmic fairness beyond a narrow focus on equality metrics. Cost/benefit analysis seems to offer a principled way to think about the issue. Do you see any weakensses in Huq's argument? Can yuo reconstruct the arguement here more precisely?

Serious crimes

Huq emphasizes negative spillover effects and externalities, but he believes these are less relevant in the case of serious crimes:

> "... to focus solely on the immediate costs and benefits of a coercive intervention and to ignore externalities ... seems a plausible approach with serious crimes, where externalities are dwarfed by immediate costs and benefits." (p. 1113)

Is this true of serious crimes? Serious crimes in some communities might have greater net spillover costs than in other communities. A wealthier family can more easily cope with the loss of an adult figure than a less wealthy family.

When negative externalities can be ignored, the same threshold should presumably be imposed for whites and blacks for pre-trial coercion, other things being equal.

Mistaken decisions

An interesting—but controversial—aspect of the Huq's costs/benefit framework is that it makes little difference between actual reoffenders and non-reoffenders. Huq' writes:

> "there is no particular reason to believe that any of these spillover costs are less if the person subject to the coercion is in fact a true rather than false positive ... what should matter is the absolute cost of using a coercive tactic against a member of a minority group, net of benefit, for all members of that racial group." (p. 1127-28)

Recall the debate betwen ProPublica and Northpointe. The question was, is predictive parity or classification parity the right measure of algorithmic fairness? Both measures are concerned with false positives and false negatives (from different angles, predictive and diagnostic). Huq rejects the common assumption behind the question.

Does welfare align with fairness?

Huq's theory purports to reconcile welfare (utility, efficiency) with racial equity (or more generally fairness). This is a strength of the theory. Huq's argument rests on two assumptions. First,

> "...most crime is intraracial, such that costs and benefits do not cross the color line by and large." (p. 1114)

Why is this assumption crucial for the argument?

Second, it assumes away circumstances in which racial equity and social efficiency come apart. These circumstances include:

> "... a policy benefits both the minority and the majority group, but the former benefit less than the latter. As a result of this gap, the extent of racial stratification increases even as the minority is benefited." (p. 1114)

> "...net gains from a policy for a majority group may exceed the net cost imposed on a minority group. Imagine, for example, a national security policy that generates significant benefits by imposing crushing burdens on a very small ethnic or religious minority." (p. 1114)

Huq thinks circumstances in which racial equity and efficiency come apart are rare. If they do come apart, Huq would prioritize fairness over welfare—at least in the absence of catastrophic general welfare losses (p. 1115)

A possible objection

Huq's proposal allows for seemingly objectionable trade-offs. For example, his consequentialist framework would mandate that a *less* stringent threshold apply for algorithm-based coercion that prevented serious crimes from occurring in black communities so long as the prevention of serious crime in black communities had greater overall net benefit. On this view, coercing blacks at higher rates than whites would be justified provided coercion had positive net benefits for both communities. Is this an intended result of the theory?

See on this point, footnote 307, citing Forman's 2017 book Locking Up Our Own showing how black politicians in the 70's and 80's believed that police coertion should be exercised more stringently in black comunities to combat rampant drug crime whose reduction would benefit black communities the most. Some use this argument to justify racial profiling—that the net beneficiaries of racial profiling are residents of high crime black neighborhoods. See Risse and Zeckhauser (2004), Racial Profiling, Philosophy and Public Affairs, 32(2): 131-170.