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, it has quite a bit to say!

Equal protection jurisprudence

Let's start with Huq's analysis. As a preliminary point, 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 focused on two main ideas: limiting the use of racial classification; and banning racialized intention. A third idea—limiting disparate racial 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

Racialized intent

Any differential treatment that is due to racial animus or discriminatory intent is clearly illegal. The use of race as a factor 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 specified contexts, for example, in college admissions for the purpose of fostering diversity; see Fisher v. University of Texas (2016).

Algorithms hardly have a racially discriminatory intent. Perhaps, the argument could be that algorithm rely on racially tainted historical data and using such algorithms counts as a racially discriminatory intent. But tainted 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)

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

Racial classification

Racial classification is usually prohibited, partly on the grounds that an epxlicit use of race as criterion for 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 can be used as part of a suspect-description. Federal Appellate courts have not objected to this practice and the US Sup Ct has declined to intervene.[2]

Huq thinks that the use of race as one of the predictors in an algorithmic classifier is aking to the use of race in suspect descriptions:

> . . . 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. . Indeed, 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??

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 (1103).

Racial disaparity

What about *disparate impact* as invoked in Title VII 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. In McClensky v. Kemp (1987), the US Supreme

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

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.

Cost/benefit analysis

If the legal framework is either obsolete or inapplicable, what else is there? Huq favors 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)

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 more far reaching 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 negative spillover effects on family life, employment and racial stigma, are likely to disproportionately affect minorities. As Huq writes:

> 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 effects, the standard for imposing coercion must be *more stringent* for blacks than for whites, other things equal.

This is a powerful argument. The emphasis on spillover costs has the merit of broadening the discussion about algorithmic fairness beyond a narrow focus on different equality metrics. Cost/benefit analysis offers a more principled way to think about the issue. But there are complications.

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)

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.

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.

Mistaken decisions

The costs/benefit framework makes little difference between actual reoffenders and non-reoffenders. Huq's utilitarian calculus is not sensitive to this difference:

> .
>
> there is noparticular 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 didsagreement was, is predictive parity or classification parity the right measure of algorithmic fairnerss? Both measures are concerned with false positives and false negatives (from different angles, predictive and diagnostic). Huq rejects the common assumption behind that debate.

A possible objection

Huq theory purports to reconcile welfare (or utility, efficiency) with racial equity (or more generally fairness). This reconciliation is a strength of the theory. But as any utilitarian theory, Huq's proposal allows for seemingly objectionable trade-offs.

For example, Huq's consequentialist framework would mandate that a *less* stringent threshold apply for algorithm-based coercion meant to prevent serious crimes in black communities so long as the prevention of serious crime in black communities has greater net benefits. On this view, coercing blacks at much higher rates than whites is justified provided coercion has positive net benefits for both communities. Is this an intended result of the theory?