Algorithmic Fairness – Structural Unfairness (cont'ed)

Marcello Di Bello - ASU - Fall 2021 - Week #9

Our goal today is to understand, by reading Zimmerman and Lee-Stronach's work[1], how the existence of background structural injustice should inform our assessment of algorithmic fairness.

Before we start, what is structural injustice?

> "By 'structurally unjust', we mean that the present rules and conventions of society systematically burden some with adverse health, employment, education, and other outcomes" (Section 2, p. 6)

[1] Proceed with Caution, *Canadian Journal of Philosophy*, forthcoming. Most quotations here are from this paper unless otherwise noted.

Wide and Narrow Procedural Justice (Sec. 1 and 2)

The standard, narrow account of procedural justice is usually formulated as follows: similar cases should be treated similarly, and different cases should be treated differently. More succinctly, *treat alike cases alike*, or *Like Cases Maxim* (LCM). To this narrow account, Zimmerman and Lee-Stronach add the following:

> "*Wide Procedural Justice*: Under nonideal conditions, procedural justice in the context of algorithmic systems requires
>
> (i) that we treat like cases alike in a way that is sufficiently sensitive to how structural injustice renders individuals and groups (dis-)similar
>
> (ii) that human decision-makers relying wholly or in part on algorithmic procedures cautiously and critically scrutinize algorithmic decision outcomes in ways that promote the aim of ameliorating substantive injustice, including substantive structural injustice." (Section 1, p. 5)

Clause (i) is a wide reading of LCM

Clause (ii) is discussed more extensively in Sec. 5

Consider a mortgage lending algorithm that assesses people credit risk. If the algorithm treats people with the same credit risk the same, it would satisfy narrow LCM.

> the lending algorithm… will treat individuals with the same credit risk equivalently, while failing to take into account different (dis-)advantageous effects of structural injustice on different individuals' credit risk…this system proceeds as if structural injustice were not a moral consideration relevant to a justifiable assessment of 'similarity' in the context of the Like Cases Maxim (Section 2, p. 8)

Zimmerman and Lee-Stronanch hold that algorithms operating under strctural injustice are substantively and procedurally unjust. What do we gain by identifying a distinctive notion of procedural injustice?

Clause (i) in Wide Procedural Justice, by contrast, requires that people with the same risk score but who are differentially subject to structural injustice be treated differently.

Zimmerman and Lee-Stronach examine two strategies for achieving algorithmic fairness – i.e. excluding protected attributes (Sec. 3) and formulating fairness criteria (Sec. 4). They find these strategies wanting:

In Sec. 3 and Sec. 4, Zimmerman and Lee-Stronanch are mostly concerned with clause (i) of Wide Procedural Justice. They show that existing approaches do not understand 'similarity' in LCM in a way that adequately takes into account structural injustice. In Section 5, they will discuss clause (ii).

> "these strategies rely on an incomplete model representation of which cases are truly 'similar' given existing structures of injustice. Securing procedural justice requires, at a minimum, that we assess the likeness of cases in a way that recognizes the moral relevance of background social structures to the decision at hand." (p. 2)

Excluding Protected Attributes (Sec. 3)

Consider first restricting the input variables on which algorithmic decisions are based, for example, excluding protected categories such as race and gender. Three problems arise.

1. *redundancy*: protected attributes correlate with unprotected ones
2. *objectionable goal*: excluding protected attributes (blind justice) upholds the status quo and perpetrates injustice
3. *obfuscation*: excluding protected attributes makes it impossible for decision-makers to take into account structural injustice (it creates a *deliberative gap*)

On the third problem, Zimmerman and Lee-Stronach write:

> "we cannot treat similar cases similarly if we do not know which individuals are truly similarly positioned, factoring in the extent to which their respective advantaged and disadvantaged social positions have been shaped by structural injustice." (Section 3, p. 12)

Fairness Criteria (Sec. 4)

Instead of excluding protected attributes, another approach for achieving algorithmic fairness is through compliance with fairness criteria: individual fairness, classification parity, predictive parity, calibration, causal and counterfactual criteria. These fairness criteria are used as constraints in the design of predictive algorithms.

We examined these different fairness criteria in earlier classes. See notes for week 5 and week 8.

This approach is not subject to the problem of redundant encoding and objectionable goal (*why not?*), but is still subject to the obfuscation problem:

> "[they] rely on insufficiently informative inputs, resulting in an *insufficiently informative model representation* of what constitutes 'similarity', and thus of what it means for an input feature to 'make a difference' for a decision. Such fairness strategies give us an *incomplete* picture of the impact of structural injustice: call such strategies 'Insufficiently Informative Input Strategies' (IIIS)." (Section 4, p. 14)

Recall that obfuscation creates a deliberative gap by which decision-makers are unable to take structural injustice into account. What should a sufficiently informative model representation of similarity look like? Presumably, it should be a model representation of similarity that takes into account structural injustice. If so, what should that model look like?

Similarly situated individuals

When are two individuals 'similar' for the purpose of an algorithmic decision? At this point, procedural justice needs to be supplemented by more substantive considerations. To quote the legal philosopher HLA Hart:

> any set of human beings will resemble each other in some respects and differ from each other in others and, until it is established what resemblaces and differences are relevant, 'Treat alike cases alike' must remain an empty form. To fill it we must know when ... cases are to be regarded as alike and what differences are relevant.[2]

[2] HLA Hart, *The Concept of Law* (2nd ed), Chapter 8, p. 149

If the notion of 'similar individuals' should be informed by our conception of substantive justice, this conception, however exactly articulated, should be aimed at removing structural injustice. Zimmerman and Lee-Stronach write:

> "any plausible conception of substantive justice must include principles for ameliorating *structures* of past and current injustice, rather than distributing benefits and burdens amongst individuals without any attention to structural advantages and disadvantages" (p. 4)

The conceptions of algorithm fairness based on fairness criteria are guided by superficial notions of similarity that are only tangentially sensitive to structural injustice.

As an exercise, think abot what kind of notion of similarity is implicit in fairness criteria such as predictive parity, classification parity or counterfactual fairness.

Two warnings

In taking into account how race and other protected characteristics are intertwined with structural injustice, Zimmerman and Lee-Stronach draw a distinction between essentialist attributions and ascriptions:

a) "*Essentialism vs. Ascription*: many input features are socially constructed and contested. Thus, it is not apt to think of 'race' as 'making a causal difference for an algorithmic prediction P'. It is more apt to think of a 'race ascription plus the social meaning of social positions and practices' as 'making a causal difference for P'." (Section 4, p. 17)

What difference does this distinction make in practice? Suppose we impose a different threshold of decision for individuals belonging to different racial groups. Could this strategy be objectionable because it essentializes race instaed of ascribing it?

Another complication comes from intersectionality:

b) "*Complexity and Intersectionality*: representing social structure-relevant input features in an isolated way is insufficiently informative if and because such a representation fails to model complex interactions between features that result in social (dis-)advantages." (Section 4, p. 18)

Fairness criteria and intersectionality

Many group fairness criteria, even causal and counterfactual criteria, struggle to account for intersectionality. For example,

How does the Proceed with Caution Approach account for intersectionality?

> "counterfactual strategies must answer the non-trivial question of *which counterfactuals matter* when we attempt to model intersectionality: what is a plausible counterfactual set of features for an individual who is a heterosexual Asian woman, or a queer Black man—and given the complex nature of intersectional (dis-)advantages, are all counterfactuals straightforwardly *comparable* with each other?" (Section 4, p. 20)

Proceeding with caution (Sec. 5)

In this section, Zimmerman and Lee-Stronach's concern is not so much with how algorithms should be designed so that they can properly model structural injustice, but rather, how humans who interact with algorithms should make decisions in light of background structural injustice.

> "If human decision-makers neglect to acknowledge and intervene in how unjust social structures affect algorithmic procedures, and the ways in which unjust algorithmic procedures affect the social world, they are liable to adopt wrongful beliefs about those subject to algorithmic systems, and to act wrongfully based on those beliefs." (Sec. 5, p. 21)

The key guiding oprinciple for human decision-makers is to avoid doxastic negligence (p. 21):

> *Doxastic Negligence*. A is doxastically negligent if A, purely on the basis of an algorithmic output concerning B, adopts a belief about what kind of treatment of B is warranted.

Human decision-makers, say judges, who make decisions on the basis of algorithmic predictionss should not be doxastically negligent. They should gather as much information as possible about the structural injustice that may affect (and confound) the algorithm prediction.

Closing the deliberative gap

Here is Zimmerman and Lee-Stronach's proposal, intended to avoid the obfuscation problem and close the deliberative gap:

> *Caution About Outputs (CAO)*: Instead of restricting inputs or relying on insufficiently informative inputs, human agents relying on algorithmic procedures should
>
> (i) aim to work with inputs that are maximally informative with respect to the impact of structural injustice, and
>
> (ii) exercise caution when basing decisions on algorithmic outputs. This may, depending on the context-sensitive features, entail suspending belief and remaining agnostic about a particular issue (e.g. a question like "does individual A merit X?"), or it may entail not taking any action for the moment. (Section 5, p. 22)

As an exercise, think about how a judge who is tasked with making decisions about preventative detention or sentencing based on predictive algorithms. How should the judge follow CAO?

Note that CAO is an expanded version of clause (ii) in Wide Procedural Justice.