Canadian Journal of Philosophy (2022), 52: 1, 6–25
doi:10.1017/can.2021.17

ARTICLE

Proceed with Caution

Annette Zimmermann[1,2]* and Chad Lee-Stronach[3]

[1]Department of Philosophy, University of York, York, United Kingdom, [2]Carr Center for Human Rights Policy, Harvard University, Cambridge, MA, USA and [3]Department of Philosophy and Religion, Northeastern University, Boston, MA, USA
*Corresponding author. Email: annette.zimmermann@york.ac.uk

Abstract

It is becoming more common that the decision-makers in private and public institutions are predictive algorithmic systems, not humans. This article argues that relying on algorithmic systems is procedurally unjust in contexts involving background conditions of structural injustice. Under such nonideal conditions, algorithmic systems, if left to their own devices, cannot meet a necessary condition of procedural justice, because they fail to provide a sufficiently nuanced model of which cases count as relevantly similar. Resolving this problem requires deliberative capacities uniquely available to human agents. After exploring the limitations of existing formal algorithmic fairness strategies, the article argues that procedural justice requires that human agents relying wholly or in part on algorithmic systems proceed with caution: by *avoiding doxastic negligence* about algorithmic outputs, by *exercising deliberative capacities* when making similarity judgments, and by *suspending belief and gathering additional information* in light of higher-order uncertainty.

Keywords: Artificial intelligence; procedural fairness; like cases maxim; structural injustice; uncertainty; causal interpretations of algorithmic fairness; doxastic negligence

Public and private sector entities are increasingly delegating decision-making to algorithmic systems to make predictions about our creditworthiness, our propensity for criminal behaviour, our access to welfare benefits and services, our prospective academic outcomes, or our expected job performance if hired, to name just a few examples. To those implementing these systems, algorithmic decision-making seems inherently impartial and objective. However, recent evidence concerning the impact of algorithmic decision-making and decision support systems shows that even when those who design and implement them have good intentions, these systems can magnify injustice, not reduce it. Much recent work on algorithmic fairness has explored this phenomenon from the point of view of *substantive* justice—often understood in terms of fair distributions of outcomes—while assuming that algorithmic systems are at least procedurally just. We question the latter assumption and argue that in contexts of pervasive structural injustice, algorithmic systems also fail a necessary condition of *procedural* justice: the Like Cases Maxim (LCM), which holds that individuals with morally equivalent sets of features should receive the same treatment. Procedures—including algorithmic procedures—cannot satisfy LCM if they fail to adequately model comparative differences between individuals differently affected by *background structures of injustice*. Human decision-makers relying on algorithmic decision-making systems thus run a significant moral risk of acting procedurally unjustly, unless they deliberate about and intervene upon algorithmic procedures in specific structural injustice-sensitive ways.

Section 1 argues for a wide interpretation of LCM that explicitly accounts for individuals' social structural contexts and that imposes doxastic and deliberative duties on decision-makers so as to

mitigate their risk of contributing to structural injustice. Section 2 introduces an example of an algorithmic decision-making procedure, which illustrates why many contemporary algorithmic systems fall short of a plausibly wide conception of procedural justice. Sections 3–4 critically examine formal algorithmic fairness strategies, arguing that 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.

Section 5 develops our positive counterproposal: we argue that our extended conception of procedural justice ("Wide Procedural Justice") requires that information about the effects of structural injustice be considered when designing algorithmic systems, and that human decision-makers deliberate about and respond to such information with caution. In decision contexts of high empirical complexity and moral risk, such as when using predictive algorithmic systems in structurally unjust social contexts, we are morally required to avoid *doxastic negligence*: that is, prematurely adopting beliefs and pursuing interventions on the basis of highly uncertain evidence obtained from algorithmic systems.[1] Section 6 concludes.

1. Procedural justice in a structurally unjust world

Structural injustice exists when institutions and social practices harm groups of individuals by creating and reifying social positions that are associated with complex advantages and disadvantages within a larger-scale framework of social relations (Young 2011, 39). Historically and to this day, structural injustice reflects, reifies, and compounds such (dis-)advantages experienced by members of socio-demographic groups defined by ascriptions of attributes like race, gender, class, disability, and sexuality.

Much of this structure consists of procedures: regularised sets of rules that make a complex world more manageable for decision-makers. Like other philosophers concerned with the problem of structural injustice, we think that tackling this problem requires, in Young's words, that we "shift from a focus [purely] on distributive patterns to procedural issues of participation in deliberation and decisionmaking" (Young 2011, 34). But doing so is not philosophically straightforward: What does procedural justice look like in a structurally unjust world?

We argue that algorithmic decision-making procedures can perpetuate structural injustice if they fail to reflect, and impede human agents' ability to critically scrutinize, relevant information in the outputs of algorithmic models, as well as information about the data underpinning such models. On our view, an algorithmic decision procedure is *procedurally* unjust to individuals subject to it *if* and *because* the procedure fails to include relevant information about the effects of current and past *substantive* structural injustices, including—but not limited to—racial and gender injustice.

Existing philosophical accounts of procedural justice typically define it by distinguishing it from *substantive* justice, which pertains to the justice of some allocation of benefits and burdens in society. Procedural justice, by contrast, pertains to rules and practices determining that allocation: procedures can be more or less just depending on whether they allow all those subject to them equal opportunities to advance their claims, to participate as equals in the contestation of decision outcomes, and to present relevant evidence, for example. Despite philosophical disagreement on the question of which normative criteria are sufficient for procedural justice, there is widespread agreement on the basic idea that *treating like cases alike* is a necessary condition for procedural

[1]We take this view to be aligned with Young's (2003, 7) argument that a more complete conception of justice requires considering "how the institutions of a society *work together* to produce outcomes that support or minimize the threat of domination […] Social justice concerns [individual] actions […] on the policies of particular institutions only secondarily, as [the latter] contribute to constituting *structures* that enable and constrain persons,," though we add to this view by emphasizing the distinctly *epistemic duties* of human agents when considering the relevance of structural injustice for identifying normative requirements for procedural justice.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

justice: that the same rules are consistently applied to everyone, and that those who share similar features—however we might define them—obtain similar decision outcomes (Aristotle 2000 V.3 and 1998 III.9, III.12; Hart 1961; Winston 1974; Dworkin 1986; Schauer 1987; Raz 1992; Rawls 2009).

> *Standard Procedural Justice*: Procedural justice requires that similar cases are treated similarly, and different cases are treated differently.
>
> ("Like Cases Maxim")

Considerations of procedural justice and substantive justice are linked in complex ways. Often, substantive justice is deemed morally weightier than procedural justice: we ultimately care more whether a decision outcome establishes a just final allocation than we care about how that outcome was brought about. This, however, does not mean that procedural justice cannot have independent value: something is lost when we fail to treat like cases alike, and when those subject to ostensibly substantively just decision lack opportunities to critically scrutinize the reasons why the outcomes reached *in fact* meet the demands of substantive justice. Thus, while procedural justice is not sufficient for substantive justice, it is still important for all-things-considered justice.[2]

While we remain agnostic in this paper about which conception of substantive justice is best all-things-considered, we *are* committed to the following claims: first, that 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; and second, that attending to the procedures by which such structures are to be ameliorated is itself a requirement of justice all-things-considered: a failure to articulate and follow just decision procedures therefore undermines all-things-considered justice. We are thus committed to the view that all-things-considered justice requires identifying and implementing a conception of procedural justice that responds adequately to prevailing background structures of substantive injustice, which is why we defend a *broader and more demanding* conception of procedural justice than other contributors to the philosophical debate have adopted. In this paper, we apply this higher-order view specifically to *algorithmic* procedures, though we think that our arguments about algorithmic procedures generalise more widely to other types of procedures as well, though they are particularly morally urgent in the context of contemporary AI.

As we shall argue, procedural justice requires more than simply treating like cases alike—it also requires explicitly *modeling* and *responding to* the extent to which current and past substantive injustices are (part of) the reason *why* some individuals (i) receive disadvantageous outcomes or (ii) are adversely affected by seemingly nondisadvantageous outcomes. We call this:

> *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, *and* (ii) that human decision-makers relying wholly or in part on algorithmic procedures cautiously and

[2]In *some* cases, procedural justice is *both necessary and sufficient* for substantive justice. Consider fair coin tosses, for which it does not make sense to think of substantively just outcomes as being just *independently* of the procedure used to reach that outcome, because the outcome of the coin toss is fair precisely because the procedure itself is fair. Rawls distinguishes two irreducible dimensions of procedural justice: "perfection" and "purity" (2009, 176, 318). Perfect procedural justice pertains not to the features of the process itself, but to its ability to secure a distribution that is fair by the lights of *procedure-independent* normative standards. By contrast, pure procedural justice pertains not to the resulting distribution, but to the features of the process itself, as in a fair coin toss. While we agree that there is a subset of pure procedures in principle, we do not view algorithmic decision procedures necessarily as part of that subset: whether an algorithmic procedure is fair *will* often depend on whether the procedure's outcomes satisfy procedure-independent normative standards of substantive justice, the latter of which include (but are obviously not limited to) the amelioration of structural injustice.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

> critically scrutinize algorithmic decision outcomes in ways that promote the aim of ameliorating substantive injustice, including substantive structural injustice.

Recent applications of artificial intelligence in public- and private-sector decision-making—such as algorithmic recidivism risk prediction in a criminal justice context (Angwin et al. 2016), the algorithmic allocation of welfare benefits and services (Brown et al. 2019), or algorithmic rankings of applicants during university admissions and hiring processes (Raghavan et al. 2020)—are paradigmatic examples of procedures that are unjust in this "wide" sense. While such systems can be statistically powerful—they are often able to yield sufficiently accurate outputs concerning *individuals* subject to algorithmic procedures—they fail to attend to how, in a nonideal context, outputs that are sufficiently accurate on an individual level can still entrench structural disadvantages linked to social group membership. Here, we do not mean to imply that the *best* or the *only* solution to structural injustice necessarily requires *algorithmic* interventions. It may often be better to instead change laws and institutions, or to reexamine the purpose of using algorithmic decision-making in a given domain at all: technology alone cannot solve the much larger problem of social inequality. Our point is that *if* human decision-makers decide to use algorithmic systems in a given domain, algorithmic procedures must be designed so as to capture less readily apparent but *justice-relevant* features shaped by structural injustice.

2. A deceptively simple example

Contemporary algorithmic systems aim to optimize for predictive accuracy. To illustrate how this works, consider

> *Automated Mortgage Lending:* A US mortgage lending company aims to maximise its expected profits when approving or denying loan applications. Although it is not permitted by law to base its decisions on the "protected attributes" of applicants, it may do so based on the estimated risk that the applicant will default on the loan ("creditworthiness"). As is increasingly common, this particular company relies on machine learning algorithms estimating an individual's credit risk based on how the features of that individual's application are associated with statistical patterns inferred from the population level and historical lending data.

Suppose this system is highly accurate: it is able to reliably predict whether or not an individual is going to repay the loan, and to recommend approval or denial on that basis. Suppose further that it delivers the same verdicts for individuals with different sets of protected attributes once we control for their credit risk. In this case, it *seems* that the algorithmic system satisfies some version of the Like Cases Maxim: if we deem credit risk to be the relevant normative property, then the treatment of individuals with equivalent credit risk are treated equivalently.

While evaluating applications this way is legally compliant, these systems face a significant moral and political challenge: they are liable to ignore and compound existing historical and structural injustices. By "historically unjust," we mean that marginalised groups in society experienced severe injustice in the past for which neither they nor their descendants received restitution. By "structurally unjust," we mean that the present rules and conventions of society systematically burden some with adverse health, employment, education, and other outcomes. Thus, according to this simplified picture, historical injustice can be the *cause* of current disadvantage, while structural injustice explains the *persistence* of this disadvantage. In any society shaped by significant systemic inequality, the quality of many individuals' foreseeable life prospects is constrained by social structures, which, following Haslanger (2016), we will understand as shared social practices that coordinate behaviour and resources and make (un-)available particular options for individuals. When deployed in this context, algorithmic systems striving for predictive accuracy will optimise for the status quo and give verdicts that, if relied on, will cement it further.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

In the case of mortgage lending in the US context, there is a well-documented history of racist historical and structural injustices by government and private actors that explain present aggregate disparities in income, health, job security, among many other metrics, that exist between Black, Latinx, and Indigenous members of society as well as members of other marginalised groups on the one hand, and white members of the population on the other (Rothstein 2017). Historical and structural injustice causally explain why ostensibly neutral features are in fact highly predictive of particular adverse outcomes. For example, consider a single mother A who lacks affordable, reliable childcare options near her home or work. Suppose that this has impeded her career, leading to various gaps in employment, to say nothing of the gender pay-gap during those times when she was stably employed. Or consider a couple B and C who live in a historically redlined district that has limited public amenities, underfunded schools, and is in proximity to a polluting highway. The social structural factors in each of these examples would predict for and explain that these respective applicants might truly possess a high credit risk—especially if these structural features will remain in place for the foreseeable future. On the basis of this prediction, it is likely that A's, B's, and C's loan applications will be denied.

Note, however, that algorithmic systems are blind to the *explanatory* features of the situation. They issue verdicts on the basis of superficial characteristics only. Crucially, they will treat the above case as being equivalent from a predictive accuracy perspective to an individual D who presents credit risk despite not being disadvantaged by an unjust social situation (or, at least, not to the same degree as A, B, and C).

Figure 1 illustrates this point using graphical causal analysis.[3] Suppose that the algorithmic system correctly treats various features of an application as being predictive of the target property (T), credit risk. The system thus gives a prediction (P) that aims to track the credit risk of the individual, and which will support a particular decision outcome for the applicant. Crucially, however, in nonideal contexts, structurally unjust background conditions (S) affect both the features that an applicant presents (e.g., zip code) that are inputs (I) to the algorithmic system and whether or not they possess the relevant target property (T) of having high credit risk. These systems do not factor in structural phenomena that causally affect individuals, effectively treating structural injustice as an irrelevant consideration when comparing otherwise equivalent applicants.

Structural injustice is thus a confounding variable that causally affects both the dependent and independent variables: credit risk and features, respectively (Herington 2020). Note, however, that if structural injustice remains stable, the model will continue to make highly accurate predictions. If, by contrast, social conditions were to be ameliorated, so that the degree of structural injustice decreases, then features such as zip code would cease to be predictive. In other words, algorithmic systems that are optimized for an unjust society can thus optimise for accuracy while ignoring these unjust background conditions.

When morally evaluating algorithmic decision-making cases of this kind, it is tempting to focus exclusively on the obvious *substantive* injustices of individuals who are wrongfully prevented (due to past and current social structures) from receiving beneficial decision outcomes, while taking for granted that algorithmic decision-making is at least procedurally just. While we agree that procedurally just decisions can yield substantively unjust outcomes, we emphasise that many algorithmic decision-making systems operating under background conditions of structural injustice are *both* substantively unjust *and* procedurally unjust. Procedural injustice, in this context, is a distinct moral phenomenon. Specifically, the lending algorithm in our example 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. Hence, this system

[3]Figure 1 represents the model's causal assumptions via a set of nodes (circles) representing the causal variables and directed edges (arrows) representing the direction of causal influence. For brevity, we are eliding formal details, including graphical methods for deconfounding (e.g., front-door and back-door criteria). For these details, see Pearl (2009).

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

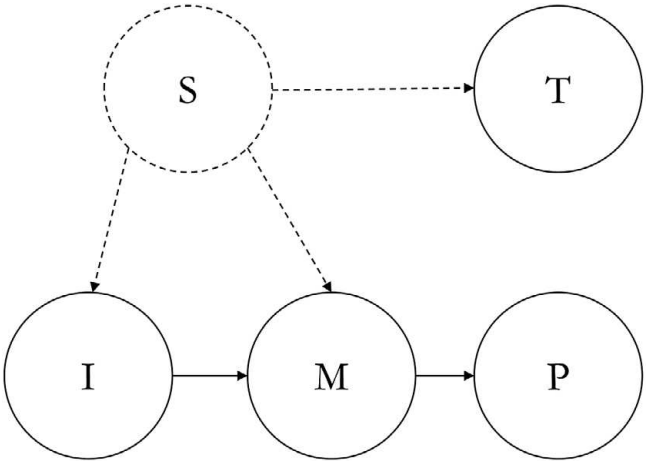

Figure 1. Structural injustice (S) is an unobserved confounding variable in an algorithmic system comprising data inputs about an individual (I), a model (M) that optimises the accuracy of its predictions (P) with respect to the target property (T).

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: where a plausibly structural injustice-sensitive conception of similarity would weigh in favour of differential treatment for (A, B, C) in comparison to (D), a structural injustice-insensitive conception of similarity will unjustly recommend equivalent treatment for differently-positioned individuals (A, B, C, D) while still seeming to treat like cases alike.

3. The *Restricted Inputs* strategy

Let us return to a seemingly simple assumption in our case above: the common idea that you cannot discriminate against an individual on the basis of feature ascriptions like race or gender if you do not *know* the relevant features of the individual. Indeed, discrimination law usually focuses on prohibiting decision-making on the basis of "protected attributes," such as gender or race.[4]

The historical roots of this normative idea reach back far and are deeply intertwined with constitutional controversy over racial segregation laws during the Reconstruction Era in the US. In the now infamous US Supreme Court decision in *Plessy v. Ferguson* (1896), in which the majority upheld the constitutionality of racist "separate but equal" laws, Justice John Marshall Harlan argued in his lone dissent that "in view of the constitution, in the eye of the law, there is in this country no superior, dominant, ruling class of citizens. […] Our constitution is color-blind, and neither knows nor tolerates classes among citizens. In respect of civil rights, all citizens are equal before the law. The humblest is the peer of the most powerful." Harlan's argument that constitutional equality requires color-blindness has significantly influenced the trajectory of progressive lawmaking and policymaking.

[4]Relevant US antidiscrimination law which articulates distinct "protected attributes" includes the Equal Pay Act (1963), the Civil Rights Act (1964), and the Americans with Disabilities Act (1990). Similarly, in the UK, the Equality Act (2010) consolidated a number of prior acts articulating protections on the basis of protected features such as (but not limited to) race and gender, including the Equal Pay Act 1970, the Race Relations Act 1976, and the Disability Discrimination Act 1995. Although there are some narrow instances in which protected features may be explicitly considered (in US constitutional law, see: *Fisher v. University of Texas*, 579 U.S. __ (2016)), our argument suggests that this class of exceptions should be significantly expanded in contexts involving algorithmic decision-making.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

This type of view may lead some to believe it is procedurally unfair for an algorithmic model to explicitly incorporate information about an individual's feature ascriptions like race and gender into its decision-making. Based on that assumption, it is natural to endorse:

> *The Restricted Inputs Strategy (RIS)*: Procedures are unfair if, and because, "protected features" like race or gender (i) *are explicitly taken into consideration*, and (ii) *make a difference* to the decision outcome. Given these assumptions, procedural justice in the context of algorithmic systems requires that we restrict permissible inputs to nonprotected features only.

Equivalently, this strategy presumes that an algorithmic decision system is procedurally fair only if it is blinded to protected features. This would require, in the case of our earlier example, that neither the data nor model mention the applicants' race, gender, or other protected attributes. This would clearly not be sufficient for procedural fairness, but proponents of this strategy would take this to be a necessary first step.

Our view is that despite the prima facie plausibility of RIS, it cannot secure procedural justice in algorithmic systems that ignore the effects of structural injustice on assessed applicants. Due to the pervasive, complex nature of structural injustice, features like an applicant's zip code will be predictive of credit risk:

> *The Redundant Encoding Problem ("It Doesn't Work")*: Restricting inputs will often not prevent algorithmic unfairness, because unrestricted inputs may function as proxies for restricted inputs.

"Blindness" does not work because under conditions of structural injustice, features like income and educational background tend to be strongly correlated with race and gender. These correlations may often be obvious—but not always. Suppose, for instance, that instead of a security asset's value, which is obviously correlated with structural injustice, one found the average temperature of the neighbourhood to be predictive of the target variable. Surely, the weather is irrelevant to structural justice! As it turns out, however, the historical impact of redlining—a racially discriminatory practice that effectively prohibited loans to Black and other marginalised communities—is visible in the average temperatures of neighbourhoods (Plumer and Popovich 2020). Today, redlined neighbourhoods are abundant in heat-radiating pavement, and lack cooling green spaces and trees, which leads to higher average temperatures compared to non-redlined areas. These nonobvious causal effects of historical injustice turn up in algorithmic models as statistical associations. Automated decisions made solely on the basis of these associations will perpetuate them.

The insufficiency of RIS is now well recognised in contemporary contributions to fair algorithmic decision-making scholarship in computer science (Dwork et al. 2012).[5] However, beyond the *technological* problem that restricting inputs does not work, and is thus not sufficient for procedural justice, there are two more fundamental *normative* worries about RIS. Consider first:

> *The Objectionable Goal Problem ("Even if it worked, it would be bad")*: The Restricted Inputs Strategy is implicitly committed to a goal which remains morally and politically objectionable as long as unjust background conditions obtain.

[5]However, this literature has focused on overcoming the problem of redundant encoding on a *technological* level by supplementing RIS with additional constraints such as ensuring that these systems have equal error rates across different combinations of protected features. We discuss these more technologically sophisticated solutions in section 4, which we argue obscure various morally relevant features of an individual's claim, and which are therefore still vulnerable to versions of the *normative* objections to RIS outlined in the remainder of this section.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

The "blind justice" view built into RIS implicitly assumes that neutral decision-making will have equal *effects*. But the opposite is true: neutrality under nonideal conditions of background injustice will likely have unequal effects which disproportionately disadvantage members of oppressed and marginalised groups, while upholding the unjust status quo (Mills 1998; Anderson 2010; Medina 2013). Empirical and historical research on the impact of colour-blind policies has shown that, despite the egalitarian intentions underpinning such policies, they lead to observable racial disparities (Alexander 2010; Bonilla-Silva 2013). Ultimately, the "blind justice" view rests on the flawed assumption that all those subject to it are starting from a reasonably just baseline. The denial of unjust background conditions *itself* is harmful not only because it may result in substantively unjust outcomes, but also because it may constitute *expressive* harm: it communicates the false message that directly experienced conditions of inequality and injustice are not real, which has a demeaning effect on those burdened by such conditions. The failure of government and other public actors in particular to publicly acknowledge as much when articulating policies adds insult to injury, and thus risks undermining citizens' conception of themselves as free and equal members of society.

Let us now turn to the final, and—in our view—the most important problem for RIS, which shows that restricting inputs (or "blinding") is not only *not sufficient* for procedural justice—it is also *not necessary*:

> *The Obfuscation Problem*: Under conditions of structural injustice, restricting inputs makes it hard or indeed impossible to determine the degree to which the ascription of sociodemographic features like race and gender *makes a difference* for outputs. This renders outputs uninformative in a morally significant way: it creates a deliberative gap for human agents interacting with the system, which undermines those agents' ability to satisfy the demands of Wide Procedural Justice.

"Protected features" like race and gender are *justice-relevant* features: it is impossible to evaluate whether a given procedure is just or not without paying explicit attention to the question of how such a procedure interacts with existing substantive structural injustices in society, many or indeed all of which track the ways in which social practices and institutions de facto position individuals differently on the basis of precisely these "protected attributes." This is a generalisation of the problem of confounding described in section 2. Due to the fact that structural injustice acts as an unobserved confounding variable, RIS do not, as a matter of *fact*, achieve their stated aims—though this is not purely a matter of technological efficacy, but also (and primarily) a *moral and political* problem. Human decision-makers ought not to design, and let their judgments be determined by the outputs of, algorithmic systems which obfuscate the degree to which outputs are shaped by structural injustice, because doing so would be wilfully ignorant. Choosing to pursue RIS, then, means choosing to risk being complicit in upholding injustice. Given that procedural justice necessarily requires—as we have argued in section 1—that we treat similar cases similarly, it is clear that blindness does not satisfy this necessary condition: 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.

In sum, our view is that procedures—including algorithmic systems—must *control* for the impact of historical and structural injustices when assessing individuals, rather than being blind to them. So far, we have shown that as long as structurally unjust background conditions obtain, rendering procedures blind to morally relevant features of individuals that are causally tied to these injustices is not *sufficient* for procedural justice, because restricting inputs allows such injustices to unduly affect and distort algorithmic procedures, and thereby to further perpetuate structural injustice. But we have also argued that restricting inputs is not *necessary* for procedural justice either: what is necessary instead is to control for the effects of background injustice in order to meet

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

the necessary condition for procedural justice to treat similar cases similarly. Controlling is not only an empirical exercise of understanding the legacy of past policies and social events; it is also a substantive moral assessment. It brings into consideration the historical and ongoing social structural factors that affect the prospects of the individual being assessed. It does so due to their moral relevance to assessing their case, not simply for their predictive value.[6]

4. *Insufficiently Informative Input* strategies

4.a Unrestricted inputs still obfuscate the impact of structural injustice

There are many technologically sophisticated alternatives to Restricted Inputs Strategies: recent formal fairness criteria include equal prediction measures, either conditional on outcomes[7] or conditional on decisions;[8] group fairness measures[9] and individual fairness measures;[10] and causal approaches.[11] Each of these approaches aims to eliminate the statistical effect of particular demographic feature ascriptions, such as race or gender. The idea is that rendering the statistical effect of these features ascriptions null amounts to ensuring that the procedures are, in a statistical sense at least, fair.

Those who aim to achieve this kind of fairness do not intervene in the social world, but in the algorithmic model. They do so via multiple stages of the model design process: for instance, data-based debiasing efforts in the form of weighing and feature selection at preprocessing with the aim of satisfying a particular set of formal fairness criteria (Bolukbasi et al. 2016; Kamiran and Calders 2012), adversarial debiasing at training time or other model-based algorithmic interventions (Zhang, Lemoine, and Mitchell 2018), and defining outcome thresholds in order to constrain the outcome set to the outcome subset that one deems unobjectionable by the lights of particular formal criteria at postprocessing.[12]

These strategies respond directly to the problems that RIS faces: there is an emerging consensus in the technical debate on fair machine learning that "blind" approaches will not achieve fairness. While all aforementioned strategies are clearly superior to RIS because they work better in terms of achieving their stated aims and because they are not implicitly committed to an objectionable moral and political goal, they are still vulnerable to the Obfuscation Problem.

Recall that Obfuscation undermines procedural justice because it hides the degree to which substantive structural injustice shapes both inputs and outputs, thus stifling the aim of treating similar cases similarly in a way that is sufficiently accurate *given a nonideal real-world social context.* One reason why a given algorithmic fairness strategy can be vulnerable to *Obfuscation* is that it might rely on insufficiently informative inputs, resulting in an *insufficiently informative model*

[6]In this way, we are offering a broader alternative to RIS than others have advocated for. For example, Corbett-Davies and Goel (2018) argue that protected attributes should be included for forward-looking reasons relating to risk assessment, which can allow these procedures to maximise aggregate social utility. Our approach allows for information relating to backward-looking considerations, such as historical injustice, to be included in the decision-making process.

[7]One example for this strategy are efforts to achieve Equality of True Positive Rates and Equality of False Negative Rates, also known as "error rate balance," "equalized odds" or "separation." For surveys of state-of-the-art formal algorithmic fairness strategies, see Barocas, Hardt, and Narayanan (2020) and Mitchell et al. (2021).

[8]For example, a criterion that requires both Equality of Positive Predictive Value and Equality of False Discovery Rate, also known as "predictive parity" or (for closely related notions) "sufficiency" or "calibration within groups."

[9]Also known as "statistical parity" or "demographic parity" (Corbett-Davies et al. 2017).

[10]This is also known as "metric fairness" (Yona and Rothblum 2018).

[11]Examples include "individual counterfactual fairness," "conditional counterfactual fairness," and "counterfactual parity" (Kusner et al. 2017; Nabi and Shpitser 2018).

[12]See Hardt, Price, and Srebro (2016) for an example of a simple postprocessing step which intervenes solely on the basis of objectionable joint statistical distributions of a predictor, a target variable, and a protected attribute, but *not* on the basis of individual features; see Dwork et al. (2018) for a strategy using transfer learning for subgroups about which we have less data with the goal of maximizing accuracy across groups.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

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).

Our central concern is that adopting IIIS increases the risk that human decision-makers will *mistakenly* think that they have achieved procedural justice after having implemented such a strategy. If we adopt false beliefs about whether we have acted (un-)justly—say, by basing our decisions on an algorithmic model that obscures important aspects of injustice—we risk upholding and further compounding structural injustice. A morally superficial modeling of similarity and difference-making therefore has important but underappreciated *epistemic* costs which matter morally and politically. After outlining why adopting a sufficiently rich notion of treating similar cases similarly is a necessary condition for procedural justice, we identify two reasons why *all* aforementioned strategies are, in principle, liable to model similarity in a way that is insufficiently informative for human decision-makers (section 4.b). We do not mean to imply that all existing *formal* fairness strategies are doomed to fail: in fact, many strategies are, in principle, compatible with amendments and constraints rendering them less morally superficial. Here, we merely explain why a less morally superficial, sufficiently informative notion of similarity *matters,* why it matters specifically *for procedural justice*, and *how much closer* a sufficiently informative algorithmic representation of similarity would get us to all-things-considered substantive justice.

Why is treating like cases alike morally and politically valuable? Failing to treat like cases alike, all else being equal, would be objectionably arbitrary (Hart 1961; Dworkin 1986; Rawls 2009), as it would prevent the creation of a system of rules in which persons' legitimate expectations are met, such that there are no unpredictable rule exceptions and right infringements. As Rawls (2009, 50–51) argues, by "secur[ing] legitimate expectations, [LCM] excludes significant kinds of injustices." One such injustice would be the arbitrary *application* of an existing *system* of rules: predictability matters not only because persons subject to a given system have legitimate expectations concerning substantive decision *outcomes* in an individual case, but also because they will have more general legitimate expectations that the system of rules will indeed be implemented in *each* case, except in cases in which specific contextual circumstances plausibly mitigate the application of a given rule.

In a society aiming to secure justice for all, arbitrary decisions require justification. While it is plausible to argue that public actors and institutions—particularly those which are directly democratically authorized—have *special obligations* not to engage in arbitrary decision-making, similar moral obligations arguably apply to private decision-making as well, such as private companies engaged in credit and mortgage lending. Just because a private actor is not democratically authorised and therefore has no *special* moral obligations of nonarbitrariness, this does not necessarily imply that she has *no* moral obligations with respect to nonarbitrariness at all: private actors play a key role in shaping social structures in contemporary democratic societies, and are not automatically exempt from important moral and political duties. Although as [other authors in this issue argue] the law allows these businesses to *suberogate*—to choose an avoidably worse option (say, by not loaning to a creditworthy individual)—our view is that arbitrary decision-making is always pro tanto wrong. While pro tanto wrongs might easily be overridden by competing moral concerns in other contexts, the presence and enduring impact of historical racial injustice in credit and mortgage lending raises the moral stakes of arbitrary lending, thereby also raising the bar for potentially overriding moral considerations (if any).

While many philosophical accounts of the role of LCM for procedural justice focus specifically on *legal* procedures—and in particular, trial and sentencing procedures—we take a significantly *broader* view so as to include any type of algorithmic or nonalgorithmic decision procedure in the private and public sector in which justice-based considerations plausibly apply. In our view, LCM construed as a moral rather than exclusively legal principle has wide applicability in a range of decision domains, including but not limited to legal procedures. Credit-granting procedures, for instance, have historically violated LCM, leading to credit denials on the basis of racial discrimination, creating structural disadvantage across generations. LCM is a minimal bulwark against this.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

Importantly, while LCM is not itself devoid of normative content, it necessarily requires additional substantive normative principles to determine its scope of application (*cf.* Westen 1982; Sunstein 1993). As Rawls argues, LCM demands that "similar cases are treated similarly, the *relevant* similarities and differences being those *identified by the existing norms*" (2009, 50–51; our emphasis), and as Hart puts it, LCM "is by itself incomplete" (1961, 159), but that "we have, in the bare notion of applying a general rule of law, the germ at least of justice" (206). Departures from the "germ of justice," *provided that they lack a plausible normative justification with reference to principles of substantive justice*, undermine the pursuit of substantive justice. What all-things-considered substantive justice—whether on an individual or structural level—would require is, of course, a question that cannot be solved algorithmically: it requires human judgment. Even adopting a nontraditional yet plausible *broad* notion of procedural justice, as we suggest here, does not free us, then, from the difficult task of determining the evaluative standards by which to ultimately judge a social state of affairs as substantively just. As Aristotle argues, "[j]ustice […] should be equal for equal persons. But equality *in what sort of things* and inequality *in what sort of things*—this should not be overlooked" (1998 III.13; our emphasis).

Recognising the central role of substantive theorising reveals the limitations of various *formal* strategies for achieving procedural justice, including those that explicitly invoke LCM. Dwork et al. (2012), for instance, formalise LCM by relying on a Lipschitz condition:

> The Lipschitz condition requires that any two individuals x,y that are at distanced (x,y) ∈ [0,1] map to distributions M(x) and M(y), respectively, such that the statistical distance between M(x) and M(y) is at most d(x,y). In other words, the distributions over outcomes observed by x and y are indistinguishable up to their distanced (x,y).

The Lipschitz condition formalizes multiple important aspects of LCM: it imposes a *consistency* requirement—for all cases, similarly situated individuals are treated similarly—and it imposes a *spacing* constraint, such that the distance between individuals at the level of inputs (as measured by some task-specific metric that approximates ground truth as closely as possible) is equivalent to the distance between outcomes for those individuals, and thus (implicitly) nonarbitrary.

Rival formal fairness strategies, by contrast, do not commit explicitly to LCM. However, in our view, any formal fairness strategy must necessarily implicitly commit to some version of LCM, though different strategies may give different responses to the question *which types of agents* (individuals, demographic subgroups, or demographic groups) ought to be compared to each other with respect to considerations of similarity.[13] Given that all formal strategies aim to achieve algorithmic procedures which block features shaped by individual or structural disadvantage from *making a difference* for outputs, all such strategies must also rely on some notion of which features *may* permissibly make a difference for outputs, and which agents thus count as relevantly (dis-) similar.

4.b Two ways of getting similarity wrong

Many fairness strategies tacitly rely on morally superficial notions of which (sets of) features are similar to other (sets of) features. First, non-restricted input strategies may explicitly include features like "race," but this alone does not tell us much about how the difference-making mechanism between structural injustice, social categories like race ascriptions, and outcomes actually works:

[13]Indeed, as Dwork et al. (2012) point out, their account is compatible with group fairness strategies (in particular, those requiring statistical parity), and in fact LCM-style individual fairness *implies* group fairness in some cases, whereas in cases in which LCM does *not* imply group fairness, additional constraints can plausibly be imposed.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

(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."

Most formal fairness strategies seem to assume that demographic features are essential categories unaffected by social norms and practices. Our position is that this view is often wrong: many demographic features—and in particular, race—are socially constructed,[14] and thus contextually and geographically contingent (*cf.* Root 2000, 632). Given this, it is important to recognize that when selecting and modeling features in an algorithmic system, we must first make a—potentially contested—choice about what the *scope* of a given feature is: who is included in the group of persons who possess that feature, and on what grounds. In other words, we *ascribe* features to individuals.

The same is true for implementing LCM once we have made such ascriptions: we *ascribe similarity and difference* to different cases *given a prior, higher-order normative judgment of who or what counts as similar enough*, for the purposes of procedural justice: as Schauer plausibly expresses this point, "identifying what is a precedent for what is about […] ascribing likeness; and it is not about discovering, locating, or unearthing likeness. Determining […] which different events or acts or questions will in spite of those differences be treated as similar, entails the question of what a decision-maker […] deems to be similar, and not about what is actually similar in some deep ontological sense" (2018, 446). If we fail to critically evaluate which feature ascriptions we should take as a given, and which ascriptions we should contest, we fall short of enacting procedural justice. We cannot opt out of making that higher-order normative judgment with respect to similarity: formal fairness measures cannot perform this task by themselves, not only because input features alone do not give algorithmic systems sufficient information for that purpose, but also because identifying suitable higher-order principles governing ascriptions of similarity requires *moral and political* deliberation.[15]

Importantly, higher-order judgments of this kind include judgments about *moral equivalence*: when to ascribe similarity to two cases involving input features which do *not* seem similar on a surface level, but which—having factored in differing impacts of structural injustice in either case—merit a similar response. Deciding to treat cases alike when background conditions of injustice have created an uneven playing field for individuals through no fault of their own, then, does not necessarily mean that one is arbitrarily treating unlike cases alike. Rather, as we shall argue in section 5, doing so can allow us to meet the demands of the moral intuitions underpinning Wide Procedural Justice.

(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.

Working with *any* plausible list of protected features, whether that list is codified in law or not, will not adequately model complex intersectional (dis-)advantages if those features are modeled in a

[14]We endorse *social constructivism about race*, the view that race is a social rather than a natural kind, which is nevertheless "real" in the sense that race ascriptions have social meaning, and in the sense that they cause people to be positioned in tangibly advantageous and disadvantageous ways in society. We remain agnostic, for the purposes of this paper, about the question of which branch of constructivism is ultimately the most plausible, including thin constructivism (Gooding-William 1998), cultural constructivism (Jeffers 2019) and political constructivism (Haslanger 2000, 2019). We also remain agnostic about the respective merits of conservationist constructivism (the view that racial categories, though socially constructed, should be preserved for the purposes of adopting policies aiming to mitigate social differences) and eliminativist constructivism (the view that racial categories should be eliminated in the long run).

[15]Here, we are assuming that such moral deliberations can be performed by human decision-makers *only*, at least until the still distant prospect of trustworthy, value-aligned Artificial General Intelligence becomes a reality.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

purely *additive* way, because intersecting disadvantages reinforce each other in a way that does not equal the sum of their isolated causal effects.[16] Consider, for example, Ana, a Black woman who is denied a loan. It is insufficient to model Ana's case by quantifying the statistical disadvantage that Black loan applicants face and adding that to the statistical disadvantage of female loan applicants. Instead, we must recognize that injustices faced specifically by Black women are relevantly different from, and irreducible to, the sum of injustices against white women and Black men. Lending decisions may be heavily influenced by employment, where intersectional disparities are particularly hard to capture via an additive approach. Jenkins presents an intuitive example: "suppose that […] women are less likely to be in paid employment than men, and Black people are less likely to be in paid employment than women. An additive approach would [suggest] that Black women are especially unlikely to be in paid employment—they are unlikely "twice over", as it were. But […] it might be that Black women are employed at higher rates than White women and higher rates than Black men, because Black woman, unlike White women, are subject to economic pressure to take on paid work, and because there are many domestic service jobs available for which Black women, unlike Black men, are considered suitable employees" (Jenkins 2019, 264f.). However, even if Black women like Ana are employed at higher rates, they may *still* be unjustly disadvantaged in credit lending decisions in comparison to other groups, and this disadvantage will not map neatly onto the ways in which employment affects lending decisions for Black men and white women: Black women may face qualitatively distinctive injustices, which are hard to model as a mere aggregation of injustices against others in this context. Importantly, even if a purely additive model of intersectional disadvantage appears accurate in some cases, such a model might be misleadingly simple in that it might not transfer well to different, rapidly evolving social contexts, and lose its explanatory power with respect to similarity.

Many existing formal fairness strategies struggle to represent intersectional and context-dependent forms of disadvantage: consider, for instance, individual fairness strategies. It is difficult for such strategies to distinguish between individuals who *seem* to be similarly socially positioned at a given moment in time, but who—due to intersecting forms of disadvantage—may experience different repercussions of receiving similar algorithmic outputs. Group fairness strategies, in turn, must answer (though often fail to address explicitly) a number of complex, context-specific questions: who counts as part of a given group? Which groups matter from the perspective of justice? And, most urgently, is the risk of unjust disadvantage of individuals who are members of *multiple* oppressed groups, and thus experience intersecting disadvantages, best modeled as a mere combination of multiple group fairness measures stacked together, as formal group fairness accounts often do?

Supporters of group fairness strategies—as well as other formal strategies—may respond that in order to adequately model intersectional disadvantages, we could simply consider more fine-grained groups (e.g., by having a separate group for Black women), and implement formal fairness measures (such as statistical parity or equalised false positive rates) *across* all groups. This would better model intersectional disadvantages than a more coarse-grained approach. However, even the more fine-grained approach may fail to achieve fairness in cases in which, due to background structures of injustice, *equalising* across fine-grained groups would not have equal *effects*. Suppose

[16]On this *causal* interpretation of the nonadditivity thesis, we are following Bright, Malinsky, and Thompson (2016). On the theoretical foundations of the nonadditivity thesis, we endorse Crenshaw's (1991) influential view. It is worth noting that work on the nonadditivity thesis has a much longer history in Black feminist thought. Consider, for instance, the Combahee River Collective's statement that "we are actively committed to struggling against racial, sexual, heterosexual, and class oppression, and see as our particular task the development of *integrated* analysis and practice based upon the fact that the major systems of oppression are *interlocking*. The synthesis of these oppressions creates the conditions of our lives" [cited in Taylor 2017, 15; our emphasis), or Audre Lorde's (1984, 138) claim that "[t]here is no such thing as a single-issue struggle because we do not live single-issue lives." For a historical overview over the development of intersectionality theory starting in the nineteenth century, see Gines (2011).

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

that Ana and Betty (a white woman) are similarly positioned loan applicants, and both Ana's and Betty's applications are denied. Ana not getting a loan may well have more disadvantageous effects in comparison to the effects of the denied loan on Betty, given that the decision result of denying Ana the loan will interact with other future decision results in other domains of Ana's life, all of which structurally disadvantage Ana in ways that are different from the structural effects on Betty.

Several *causal* algorithmic fairness strategies have recently attempted to address the problem of modeling intersectionality directly (*cf.* Bright, Malinsky, and Thompson 2016; Yang, Loftus, and Stoyanovich 2020). Foulds et al. (2019), for instance, articulate an "intuitive intersectional definition of fairness: regardless of the *combination of protected attributes*, the probabilities of the outcomes will be similar" (our emphasis). Note, however, that such strategies still face the morally and politically complex question of which combination of intersecting features counts as morally "similar"—with respect to the extent of structural injustice experienced by those who have that combination of features—to which *other* combination of features, and which way of *quantifying* similarity might adequately model intersectional disadvantages, including the real social effects of algorithmic outputs on individuals and groups: this kind of similarity judgment is needed in order to determine which members of which groups should, as a matter of justice, receive similar algorithmic *scores*. Similarly, *counterfactual* strategies must answer the nontrivial 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? The answer to these questions will require not only sophisticated formalization, but also (and indeed, primarily) nuanced engagement with heterogeneous lived experiences and persistent disagreement.

5. How to proceed with caution

5.a Getting similarity right would not by itself achieve procedural justice

Even if Insufficiently Informative Input Strategies did not model similarity in a misleadingly simple way, they would still not meet the demands of procedural justice under nonideal conditions, because procedures cannot be viewed in isolation from the social context in which they are implemented. An algorithmic procedure is *but one part* of a larger decision procedure involving dynamic interactions between technological models, human agents, and the social world. 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. The moral duty to avoid replicating and amplifying background structures of injustice thus has an important *epistemic* component. When engaging with information from algorithmic outputs, we must avoid jumping to conclusions, lest we commit:

> *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.

By "purely on the basis of an algorithmic output," we mean "without any additional deliberation about whether that output is in fact decisive for the actions we ought to take." Note that, on our view, even if A adopts *morally right* beliefs about what treatment of B is warranted, A can still be doxastically negligent if A fails to engage in a process of deliberation on whether the algorithmic model obfuscates any justice-relevant information, whether the algorithmic procedure *in fact* treats like cases alike, and whether *more information* about B, the social structures that affect B, and the availability of means (if any) for intervening in the structures causing disadvantage must be gathered.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

This point can be tied to recent accounts of moral encroachment (Basu 2019; Bolinger 2020), the view that:

> *Moral Encroachment*: Whether an agent knows some proposition depends in part on the moral stakes of the situation.

If we accept the claim that moral norms require justified belief or even knowledge that one is not doing wrong, then we must also accept the claim that human decision-makers should not rely uncritically on algorithmic systems when there are risks of compounding structural injustices: given that the question of whether their action will ameliorate or compound structural injustice is morally *high-stakes*—both for human decision-makers and decision subjects—this raises the bar for what kinds of beliefs human decision-makers may justifiably adopt based on particular types of evidence. It follows that in our earlier lending example, even if predictive outputs concerning credit risk are highly statistically accurate, human decision-makers relying on such outputs nevertheless may not —as a matter of justice—*negligently adopt beliefs* about individuals' creditworthiness purely based on such outputs, nor may they *act upon* such beliefs (e.g., by denying loans) without sufficient prior deliberation about the attendant moral risks.

Note, however, that whether or not one accepts Moral Encroachment is not decisive for whether one can accept our argument concerning the duty to avoid Doxastic Negligence. One need not endorse Moral Encroachment in order to accept the more general normative principle that in high-stakes decision scenarios involving incomplete information, human decision-makers ought to gather more information before acting—and as our earlier argument emphasized, many seemingly straightforward algorithmic decision scenarios that ostensibly provide human decision-makers with sufficient information are, in fact, insufficiently informative with respect to the impact of structural injustice on algorithmic outputs. Within the causal analysis of our earlier lending example, critical deliberation and information-gathering would help deconfound the often-overlooked relationship between structural injustice, the individual, and the algorithmic system. Importantly, estimating the causal effect of structural injustice would (i) allow human decision-makers to control for it, (ii) estimate effects of potential justice-oriented interventions, and (iii) avoid actions which risk perpetuating structural injustice, and instead take actions that ameliorate it.[17]

5.b Closing the deliberative gap

As we have argued, anyone aiming to interrogate the assumptions encoded in an algorithmic procedure must maintain a healthy dose of doubt about the veracity of the representation of complex social facts in the models underpinning algorithmic procedures. This type of doubt can be understood in terms of *higher-order uncertainty*, where the decision-maker's own uncertainty about an uncertain set of facts is relevant to the decision problem. Avoiding doxastic negligence means deliberately recognising and leaving room for such uncertainty when evaluating which actions are morally and politically justified in light of particular algorithmic classifications. Beyond a more demanding, structural injustice-sensitive notion of LCM, Wide Procedural Justice thus requires:

> *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

[17]On deconfounding and estimating causal effect, see Pearl (2009, 65–106).

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

particular issue (e.g., a question like "does individual A merit X?"), or it may entail not taking *any* action for the moment.

Sometimes, the two necessary and jointly sufficient components of procedural justice as we conceive of it—adherence to a plausibly rich version of LCM on the one hand, and the moral and political duty to avoid doxastic negligence by proceeding with caution—may conflict. Consider cases in which *substantive* justice clearly requires that we treat descriptively *dissimilar* cases similarly in order to bring about more equitable outcomes.[18] In such cases, the value of responding to new information with caution will usually outweigh the value of strict adherence to LCM. CAO therefore also entails the duty to deliberate critically on the likely impact of structural injustice on algorithmic outputs when making ascriptive judgments about *which* cases we ought to consider (dis-)similar, and *why*. In light of more information, it may turn out that we have strong moral reasons to ascribe similarity to a broader range of cases than we currently do, or indeed to explicitly treat unlike cases alike if doing so would get us closer to a plausible ideal of substantive justice. According to Dworkin's (1986, 219) account of *political integrity*, deviations from LCM are justified from the perspective of procedural justice iff they are consistent with a higher-order theory of justice. While articulating a complete theory of which specific substantive ideals of justice should guide us in our assessment of whether a departure from LCM would contribute to political integrity would be beyond our scope, we endorse Dworkin's general view that departures from the LCM component of procedural justice may be desirable under specific circumstances.[19]

Skeptics of Wide Procedural Justice who disagree with including CAO as a necessary component of procedural justice can still endorse all *substantive conclusions* of our account, while rejecting the view that moral requirements like (i) a "thick," structural injustice-sensitive interpretation of LCM and (ii) CAO can all be unified under any single procedural justice framework. We leave open the possibility that normative principles *independent* of procedural justice could plausibly justify CAO. The reason why we include CAO as a necessary component of procedural justice in the context of algorithmic systems is that we define "procedures" broadly as the full decision sequence which necessarily includes some algorithmic procedure and some human response, rather than defining it narrowly as just the algorithmic procedure. For this reason, we view human deliberation as an appropriate object of procedural justice constraints.

To illustrate the CAO component of Wide Procedural Justice more concretely, let us return to our earlier lending example (section 2). Supposing that the algorithmic system classifies an individual B as "high risk of default," what should a human decision-maker (A) using that system do? Should A necessarily deny B's application? Although it may seem to conveniently allow A to shirk responsibility for error ("computer says no"), the algorithmic prediction does not *in fact* tell A what to do, especially when it comes to assessing B in comparison to another individual C who seems superficially similar to B, such that an algorithmic model recommends the same outcome for B and C. There is a wide deliberative gap between classification and procedurally just decisions.[20] Closing that gap, and deciding whether B should get the same rate as C, will require asking questions of the following kind:

[18]Schauer (2018, 448) makes a similar argument: "consider, for example, the statement […] "all men are created equal." […T]he statement […] was made against the background of the numerous ways in which people are not […] empirically equal. Some are smarter, nicer, fitter, […] and the authors of the Declaration plainly knew that. But the ascriptive dimension of the statement is that people should be treated as equal not because they are equal in a descriptive sense, but despite the fact that they are not."

[19]Our view differs from Dworkin's view, however, in that Dworkin allows for departures from LCM only on the basis of axioms of a pure theory of justice, whereas our account allows for departures form LCM on the basis of pragmatic, contextual, and thus nonaxiomatic reasons. For a similar point, see Rawls (2009, 51).

[20]This claim is compatible with the claim that deliberation alone will not *guarantee* substantively just outcomes, since—as we have argued in section 1—procedural justice is not sufficient for substantive justice.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

1. Has A considered the *full* range of available options (e.g., denial, approval, different loan rates or conditions, social services)?

   In simplistic decision models, available options are predefined. In more realistic cases, one must discover the options available by querying the causal structure of the situation, identifying possible interventions that might bring about more just outcomes. These interventions will be outside the scope of predictive algorithmic systems, since they address counterfactual scenarios that humans are capable of recognising, but which are not reflected in the system's data.[21]

2. How much is at stake for the decision subject (i.e., loan applicants B and C)? Will an adverse decision outcome impact B and C in the same way, or will structural injustice lead to vastly different effects for B and C?

   If A is reasonably certain that structural injustice disadvantages B in a way that makes B *relevantly different* from C, by the lights of some plausible conception of "creditworthiness," then B should get a different, more lenient rate than C.

3. How much is at stake for decision-maker A: Might A have morally relevant biases which impair their ability to know what justice would require in this context? How much is A's judgment about B being actually similar to C affected by higher-order uncertainty? Can A reasonably adopt any belief about B, or is A morally required to suspend judgment for the time being?

   If A is unsure about the extent to which structural injustice has disadvantaged B in comparison to C, A should suspend—at least temporarily—belief about B, and adopt a *blanket rule* that gives B the benefit of the doubt, the latter of which counteracts historically unjust negative assumptions about "*people like B* defaulting on loans." While suspending belief, A ought to gather additional information, while weighting the moral value of such information against potential morally significant costs of information-gathering like privacy breaches.

5.c How much caution?

Given that background structures of injustice are ubiquitous, and thus shape *many* or possibly even *all* contemporary decision scenarios, one may worry that our view prescribes an objectionably high degree of caution in an objectionably high number of cases:

> *The Caution Paralysis Worry*: CAO seems to imply that we should approach the overwhelming majority of algorithmic outputs with a high degree of caution. A duty to keep deliberating, and to suspend belief in cases involving insufficiently informative inputs, may stifle rather than support our ability to make decisions efficiently.

However, under conditions of entrenched structural injustice, the disvalue of procedural injustice at a massive scale, enabled by the increasingly widespread use of algorithmic systems, far outweighs the value of efficient decision-making. It is morally *worse* to jump to conclusions than it is to suspend belief, if one grants the—in our view, plausible—assumption that under such conditions, human decision-makers are more likely to have implicit or explicit biases that align with background injustice than they are to have biases that oppose it.[22] If decision-makers have those biases,

[21]This is a standard conceptual point made by proponents of causal modelling. For an application, see Prosperi et al. 2020.

[22]As philosophers and social scientists working on phenomena like adaptive preferences and double consciousness have persuasively shown, such injustice-sustaining biases can also adversely affect deliberations by those *disadvantaged* by structural injustice.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

jumping to conclusions—that is, committing doxastic negligence about algorithmic outputs—is likely to sustain background structures of injustice, in which case the mere fact that the decisions which sustained that state of affairs were made efficiently has barely any moral and political value.

Relatedly, one may worry that it is *undue caution* that may ultimately sustain injustice if and when the prescription to suspend belief is taken to mean *not taking any action* based on a given algorithmic output. Not taking any action, CAO skeptics may reasonably argue, does not mean that one has not made a decision: not taking any action necessarily requires choosing one option (nonintervention) over another (intervention)—or, in the context of our earlier lending case, suspending a loan application decision (nonintervention) rather than approving or denying the loan (intervention). Assuming that under background conditions of structural injustice, nonintervention will typically sustain the unjust status quo, it seems that CAO may ultimately prescribe rather than prohibit structural injustice-entrenching actions on the part of human decision-makers.

But of course, suspending belief does not necessarily mean nonintervention: Caution about Outputs does permit intervention, as long as decisions to intervene are subject to ongoing critical scrutiny and potential revision in light of new information about justice-relevant features. For example, CAO-compliant interventions could involve altering decision rules in order to approve more loans for members of structurally disadvantaged groups, or (at the very least) treating decisions to deny loans to members of such groups as highly revisable in light of mitigating contextual factors. We can avoid doxastic negligence by explicitly recognizing opportunities to *change our minds* about what justice in fact requires when we evaluate decision procedures and make judgments about which treatments count as fair.

6. Conclusion

Automating decision-making processes in public and private institutions tends to hide important moral and empirical complexities beneath a veneer of impartiality and objectivity. In particular, algorithmic systems in many contemporary settings are subject to confounding by structural injustice, leading to algorithmic predictions that further entrench rather than ameliorate unjust background conditions. Although various formal strategies exist for making such procedural systems more "fair," they are liable to obfuscate or oversimplify the influence of structural injustice: they fail to make salient the extent to which current and past injustices are (part of) the reason *why* some individuals receive disadvantageous algorithmic outcomes, or are adversely affected by seemingly nondisadvantageous outcomes. This results in a dangerously incomplete picture of what it would take to meet a key requirement of procedural justice: determining which cases are relevantly similar and treating similar cases similarly on the basis of that determination.

Meanwhile, algorithmic systems, as well as formal interventions into such systems, often endow algorithmic outputs with a false air of certainty: there is a morally and politically urgent risk that human agents relying on such systems will assume that justice has been done as soon as the algorithmic system delivers an output. But as we have argued, procedural justice in this context requires that we base our human response to algorithmic outputs on a deliberate acknowledgement of how much uncertainty remains in a given decision scenario: how much justice-relevant information is missing, which cases should truly count as similar given background conditions of injustice, and which social complexities are camouflaged by ostensibly fair algorithmic models. In light of incomplete information, the act of suspending belief—of giving decision subjects the benefit of the doubt when the full impact of structural injustice on them is unclear—is necessary for meeting the demands of procedural injustice, and it contributes to the larger-scale moral and political aim of ameliorating structurally unjust conditions.

Navigating the moral risks involved in deploying predictive algorithmic systems, then, requires moral deliberation and doxastic restraint uniquely available to human agents. For the foreseeable future at least, as our public institutions and private-sector decision processes increasingly rely on algorithmic systems, justice requires that we proceed with caution.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

Acknowledgments The authors wish to thank an anonymous reviewer for the *Canadian Journal of Philosophy* for helpful comments; David Danks, Seth Lazar, Rob Reich, and Kate Vredenburgh for extensive correspondence concerning earlier versions of this paper; and the other contributors to this special issue, as well as workshop audiences at the McCoy Family Center for Ethics in Society at Stanford University (2020) and the Department of Philosophy at the University of York (2020), for further discussion and feedback.

Annette Zimmermann is a permanent Lecturer in Philosophy at the University of York, and a Technology and Human Rights Fellow (2020–2022) at Harvard University. They conducted postdoctoral research at Princeton University (2018–2020) and doctoral research at the University of Oxford (2014–2018).

Chad Lee-Stronach is a tenure-track assistant professor at the Department of Philosophy and Religion, Northeastern University. He received his PhD in philosophy from the Australian National University in 2019 and was a postdoctoral fellow at the Center for Ethics in Society, Stanford University (2019–2020).

References

Alexander, Michelle. 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New York: New Press.

Anderson, Elizabeth. 2010. *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." 2016. *ProPublica* (May 23). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aristotle. 2000. *Aristotle: Nicomachean Ethics*. In *Cambridge Texts in the History of Philosophy*, edited by Roger Crisp. Cambridge: Cambridge University Press.

Aristotle. 1998. *Aristotle: Politics*. Translated by C. D. C. Reeve. Indianapolis, IN: Hackett.

Barocas, Solon, Moritz Hardt, Arvind Narayanan. 2020. *Fairness and Machine Learning: Limitations and Opportunities*. https://fairmlbook.org/.

Basu, Rima. "Radical Moral Encroachment: The Moral Stakes of Racist Beliefs," *Philosophical Issues* 29, no. 1 (2019), 9–23.

Bolinger, Renée Jorgensen. 2020. "Varieties of Moral Encroachment." *Philosophical Perspectives* 34 (1): 5–26.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *Advances in Neural Information Processing Systems*: 4349–57.

Bonilla-Silva, Eduardo. 2013. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States*, 4th ed. Lanham, MD: Rowman & Littlefield.

Bright, Liam Kofi, Daniel Malinsky, and Morgan Thompson. 2016. "Causally Interpreting Intersectionality Theory," *Philosophy of Science* 83 (1): 60–81.

Brown, Anna, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*: paper no. 41, 1–12.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." arXiv:1701.08230.

Corbett-Davies, Sam, and Sharad Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." arXiv:1808.00023v2.

Crenshaw, Kimberlé. 1991. "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color," *Stanford Law Review* 43 (6): 1241–99.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. "Fairness through Awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*: 214–26.

Dwork, Cynthia, Nicole Immorlica, Adam Tauman Kalai, Max Leiserson, Sorelle Friedler, and Christo Wilson. 2018. "Decoupled Classifiers for Group-Fair and Efficient Machine Learning." *Proceedings of Machine Learning Research* 81: 1–15.

Dworkin, Ronald. *Law's Empire*. 1986. Cambridge, MA: Belknap Press.

Fisher v. University of Texas 579, U.S. __ (2016).

Foulds, James R., Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2019. "An Intersectional Definition of Fairness." arXiv: 1807.08362v3 [cs.LG].

Gines, Kathryn T. 2011. "Black Feminism and Intersectional Analyses: A Defence of Intersectionality," *Philosophy Today* 55: 275–84.

Gooding-Williams, Robert. 1998. "Race, Multiculturalism and Democracy," *Constellations* 5 (1): 18–41.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." arXiv:1610.02413v1 [cs.LG].

Hart, H. L. A. *The Concept of Law*. 1961. Oxford: Oxford University Press.

https://doi.org/10.1017/can.2021.17 Published online by Cambridge University Press

Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34: 31–55.
Haslanger, Sally. 2016. "What Is a (Social) Structural Explanation?" *Philosophical Studies* 173 (1): 125–27.
Haslanger, Sally. 2019. "Tracing the Sociopolitical Reality of Race." In *What Is Race? Four Philosophical Views*, edited by Joshua Glasgow, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer. New York: Oxford University Press.
Herington, Jon. 2020. "Measuring Fairness in an Unfair World," *AIES 2020 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*: 286–92.
Jeffers, Chike. 2019. "Cultural Constructionism." In *What Is Race? Four Philosophical Views*, edited by Joshua Glasgow, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer. New York: Oxford University Press.
Jenkins, Katharine. 2019. "Conferralism and Intersectionality: A Response to Ásta's *Categories We Live By*." *Journal of Social Ontology* 5 (2): 261–72.
Kamiran, Faisal. and Toon Calders. 2012. "Data Preprocessing Techniques for Classification without Discrimination." *Knowledge and Information Systems* 33: 1–33.
Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual Fairness." *Advances in Neural Information Processing Systems (NeurIPS)*.
Lorde, Audre. 1984. *Sister Outsider*. Trumansburg, NY: Crossing Press.
Medina, José. 2013. "Color Blindness, Meta-Ignorance, and the Racial Imagination." *Critical Philosophy of Race* 1 (1): 38–67.
Mills, Charles W. 1998. *Blackness Visible: Essays on Philosophy and Race*. Ithaca, NY: Cornell University Press.
Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. "Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application* 8: 141–63.
Nabi, Razieh, and Ilya Shpitser. 2018. "Fair Inference on Outcomes." *Proceedings of the AAAI Conference on Artificial Intelligence*: 1931–40.
Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press.
Plessy v. Ferguson, 163 U.S. 537 (1896).
Plumer, Brad, and Nadja Popovich. 2020. "How Decades of Racist Housing Policy Left Neighbourhoods Sweltering." *New York Times*, August 8.
Prosperi, Mattia, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. "Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare." *Nature Machine Intelligence* 2 (7): 369–75.
Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 469–81.
Rawls, John. 2009. *A Theory of Justice*, rev. ed. Cambridge, MA: Harvard University Press.
Raz, Joseph. 1992. "The Relevance of Coherence." *Boston University Law Review* 72 (2): 273–324.
Root, Michael. 2000. "How We Divide the World." *Philosophy of Science* 67 (supp.): 628–39.
Rothstein, Richard. 2017. *The Color of Law*. New York: Liveright.
Schauer, Frederick. 1987. "Precedent." *Stanford Law Review* 39 (3): 571–605.
Schauer, Frederick. 2018. "On Treating Unlike Cases Alike." *Constitutional Commentary* 34: 437–50.
Sunstein, Cass R. 1993. "On Analogical Reasoning." *Harvard Law Review* 106 (3): 741–91.
Taylor, Keeanga-Yamahtta, ed. 2017. *How We Get Free: Black Feminism and the Combahee River Collective*, Chicago: Haymarket.
Westen, Peter. 1982. "The Empty Idea of Equality." *Harvard Law Review* 95 (3): 537–96.
Winston, Kenneth. 1974. "On Treating Like Cases Alike." *California Law Review* 62 (1): 1–39.
Yang, Ke, Joshua R. Loftus, and Julia Stoyanovich. 2020. "Causal Intersectionality for Fair Ranking." arXiv:2006.08688v1[cs.LG].
Yona, Gal, and Guy Rothblum. 2018. "Probably Approximately Metric-Fair Learning." *Proceedings of Machine Learning Research* 80: 5680–88.
Young, Iris M. 2003. "Political Responsibility and Structural Injustice." *The Lindley Lecture*. University of Kansas.
Young, Iris M. *Justice and the Politics of Difference*, rev. ed. 2011. Princeton, NJ: Princeton University Press.
Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. "Mitigating Unwanted Biases with Adversarial Learning." arXiv:1801.07593v1 [cs.LG].

Cite this article: Zimmermann, A. and Lee-Stronach, C. 2022. Proceed with Caution. *Canadian Journal of Philosophy* 52: 6–25, doi:10.1017/can.2021.17