Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy

Atoosa Kasirzadeh
University of Edinburgh, UK
atoosa.kasirzadeh@ed.ac.uk

ABSTRACT

Data-driven predictive algorithms are widely used to automate and guide high-stake decision making such as bail and parole recommendation, medical resource distribution, and mortgage allocation. Nevertheless, harmful outcomes biased against vulnerable groups have been reported. The growing research field known as 'algorithmic fairness' aims to mitigate these harmful biases. Its primary methodology consists in proposing mathematical metrics to address the social harms resulting from an algorithm's biased outputs. The metrics are typically motivated by – or substantively rooted in – ideals of distributive justice, as formulated by political and legal philosophers. The perspectives of feminist political philosophers on social justice, by contrast, have been largely neglected. Some feminist philosophers have criticized the local scope of the paradigm of distributive justice and have proposed corrective amendments to surmount its limitations. The present paper brings some key insights of feminist political philosophy to algorithmic fairness. The paper has three goals. First, I show that algorithmic fairness does not accommodate structural injustices in its current scope. Second, I defend the relevance of *structural injustices* – as pioneered in the contemporary philosophical literature by Iris Marion Young – to algorithmic fairness. Third, I take some steps in developing the paradigm of 'responsible algorithmic fairness' to correct for errors in the current scope and implementation of algorithmic fairness. I close by some reflections of directions for future research.

CCS CONCEPTS

• **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**=0.

KEYWORDS

Algorithmic Bias; Algorithmic Fairness; Algorithmic Justice; Structural Injustice; Distributive Justice; Feminist Philosophy; Political Philosophy; Responsibility; Ethics of Artificial Intelligence; Ethical Machine Learning

ACM Reference Format:
Atoosa Kasirzadeh. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://10.1145/3514094.3534188

This work is licensed under a Creative Commons Attribution International 4.0 License.

AIES'22, August 1–3, 2022, Oxford, United Kingdom
© 2022 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-9247-1/22/08.
https://10.1145/3514094.3534188

1 INTRODUCTION

Consequential social and institutional decisions are increasingly guided and automated by artificial intelligence, especially by resorting to data-driven and predictive machine learning algorithms. These algorithms are extensively employed to inform high-stakes decision making in a range of contexts, including the allocation of employment opportunities and the provision of medical care. Corresponding to the increased utilization of data-driven predictive machine learning in diverse social fields, the concern that the algorithms may exacerbate wrongful discrimination and biases has prompted increased scrutiny.

The attributed bias potential becomes especially salient in the treatment of particular social groups on the basis of sensitive characteristics, like race and gender. For example, it has been demonstrated that facial recognition algorithms perform less accurately on non-white women [8, 34]. Moreover, recidivism algorithms are claimed to be among the most systematically biased, as they arguably accord higher risk scores to non-white individuals, for instance in the US criminal justice system [1]. Or, it has been shown that language models encode discriminatory language and harmful social stereotypes against already-vulnerable groups [9, 49]. The insidiousness of such biases has perpetuated wrongful moral and social harms –– for example, by ensuring that Black Americans do not receive just and fair access to scarce medical resources [43].

By means of redress, a recent explosion of research in the rapidly growing field of 'algorithmic fairness' has burgeoned (for a comprehensive introduction to algorithmic fairness, see Barocas, Hardt, and Narayanan [2]; for useful reviews of algorithmic fairness tools and methods, see Corbett-Davies and Goel [12], Mehrabi et al. [39], Chouldechova and Roth [10], and Mitchel et al. [41]). The algorithmic fairness literature primarily seeks to mathematically formulate criteria of social justice with the avowed objective of remedying or preventing wrongful harms resulting from biased algorithmic outputs. These mathematical characterizations of social justice are then used at the level of data preparation, model learning or post-processing to improve the potential unfair biases in machine learning systems. Since social justice is the special concern of moral, legal, and political philosophy, it is warranted for researchers on algorithmic fairness to take their bearings from philosophical accounts of social justice, reflecting the best ethical, legal, or political theory thinking of the day.

To date, the primary focus of philosophical investigations of algorithmic fairness research has been rooted, in one way or another, in accounts of distributive justice. With a few exceptions outlined below, the philosophical basis of algorithmic fairness research undertaken so far has primarily aimed to render precise our intuitions about distributive justice, in particular by resolving a local resource allocation problem in order to bring about an unbiased distribution

of the material or computational goods discriminated against by algorithmic outputs.

While distributive justice – understood in local terms – is significant to the completely determined desiderata of algorithmic fairness, the exclusive focus on distributive accounts of justice invidiously deals with a single, narrowly construed aspect of justice. John Rawls who is one of the most influential contemporary political theorists and philosophers of distributive justice himself acknowledges that distributive justice addresses only a single instance of the application of the concept of justice [45, p.8]. Accordingly, the narrow scope of distributive justice carries over to algorithmic fairness.

The crucial point is that some problems of injustice may not have local distributive solutions, either because they are not the kind of problem to which a distributive solution is adequate, or because it is unclear how to accommodate the injustice issue in terms of a distribution problem within a local scope, or because it is impossible to compute one. It follows that a purely distributive justice account of algorithmic fairness runs the risk of material and social inadequacy and would become devoid of social or moral significance. This predicament raises the question of how algorithmic fairness can receive a morally and politically significant formulation. I propose to return once more to political philosophy in order to explore a constructive critique of the paradigm of distributive justice (conceived locally) with the objective of providing the necessary emendations of algorithmic fairness.

Below, I take some steps in developing the paradigm of 'responsible algorithmic fairness' designed to correct the shortcomings in the current paradigm of algorithmic fairness. In Section 2, I briefly introduce the mathematical metrics of algorithmic fairness and argue for their conceptual commitment to account of local distributive justice. Drawing upon the works of the prominent political theorist and philosopher Iris Marion Young, I show in Section 3 that the limited focus of algorithmic fairness on questions of distribution might not accommodate a crucial dimension of social justice, namely *structural injustice*. I examine in particular the reasons why structural injustice should be included in research on algorithmic fairness, and I take some steps in characterizing how the inclusion can be implemented.

My examination of structural injustice reveals that seeking one-off static solutions in terms of fairness metrics is inadequate, because narrow scope is susceptible to 'ethics washing' [4, 32, 48], insofar as making minor mathematical changes to algorithmic outcomes is supposed to be a putatively valuable ethical or legal solution to certain algorithmic decision-making problems. I argue that this problem requires widening the scope of algorithmic fairness to include power relations, social dynamics and actors and structures which are among the main sources of the emergence and persistence of social injustices relevant to algorithmic systems. In Section 4, I take some steps in developing a correspondingly broad-scope conception of social justice as the basis for designing just algorithmic ecosystems, receptive to both distributional and structural sources of social injustice. The basic idea is that some notions of responsibility can help to remove structural injustices that are relevant to fair algorithmic systems. In particular, I argue that Young's account of forward-looking responsibility is to be reflected and embedded into the paradigm of algorithmic fairness. The point of responsible algorithmic fairness is to build, in one way or another, social power relations and dynamics into our accounts of fair algorithmic predictions or decisions. Section 5 presents six positive practical consequences of adopting my proposed paradigm of responsible algorithmic fairness. Finally, in Section 6 I conclude the paper.

2 ALGORITHMIC FAIRNESS: DISTRIBUTIVE APPROACH

What is fair and how should we design machine learning systems for fair predictions or decisions? This question has been posed n reaction to the observation that many instances of machine learning applications have given rise to unfairness or bias in high-stakes decision domains. This observation has motivated various systematic efforts for formalizing fairness. These efforts range from measuring algorithmic outputs in relation to some fairness criteria such as statistical or demographic parity [15], equal opportunity [25], counterfactual comparisons between members of protected and unprotected groups [35] – to name just a few examples. Let us review some of these mathematical formalizations by focusing on an example.

Few computational decision systems have given rise to as intense a discussion as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Central to COMPAS is a recidivism risk-assessment algorithm for judging pretrial decisions concerning bail or parole. The risk-assessment algorithm predicts the likelihood that pretrial defendants will commit a violent crime in the near future or fail to appear in court. Risk-assessment decision systems similar to COMPAS are increasingly utilized in parts of the United States specifically to reduce wrongful discrimination in the criminal justice systems.[1]

The 2016 ProPublica analysis of the use of COMPAS in criminal justice in Broward County, Florida, claims that the algorithm's predictive results exhibit patterns of unjust racial bias [1, 37], suggesting that the performance of the COMPAS model unjustifiably varies relative to diverse racial groups. In particular, Angwin et al. [1] showed that, among the defendants who ultimately did not reoffend, Blacks were labelled 'high risk' by the algorithm approximately twice as frequently as whites. By contrast, among defendants who ultimately did reoffend, whites were more than one and a half times likelier than Blacks to be labelled 'low-risk' by the algorithm. Angwin et al.'s crucial observation is that the algorithm yields unjust racial bias because it manifests asymmetric predictive accuracy: it does not attribute the same false positive and false negative rates, on average, across the partition by race.

ProPublica's claim has been met with several objections. These query the specific statistical criterion for assessing the uniformity of false positive and false negative rates across the partition. The opponents defend an alternative criterion as the best metric for measuring unfair algorithmic allocations. Notably, Flores et al. [19] defend the well-calibration criterion as the proper metric for measuring harmful algorithmic bias. Flores et al. [19] reject the ProPublica analysis on the ground that for each possible COMPAS risk score, the percentage of blacks assigned that risk score who recidivated was approximately the same as the percentage of whites

[1] For a systematic examination of the COMPAS risk assessment tool, see Berk et al. [3].

assigned the same risk score who recidivated, meaning that the algorithm was in that sense equally well-calibrated. That is, for each possible risk score, the percentage of defendants assigned a risk score class who committed a crime was statistically the same. Dieterich et al. [13] claim that the predictions of COMPAS are not racially biased because the predictions were equally accurate, on average, across the partition.

Fundamentally, the counterproposals raise the question of which statistical criteria adequately and conclusively assess the algorithm's predictive performance (and how they are justified, not just in mathematical terms but in a substantive moral or philosophical sense). COMPAS is among several algorithms that (arguably) evince the exacerbation of unfairness and injustice through algorithmic decision making. Another recent case for 'The Secret Bias Hidden in Mortgage-Approval Algorithms' revealed that, in the United States, loans to non-white applicants were 40% to 80% more likely to be denied relative to the white counterpart class [38]. In metropolitan areas especially, the disparity exceeded 250%. For a collection of examples of unfair algorithmic biases, see Barocas, Hardt, and Narayanan [2].

When viewed from a conceptual perspective, most mathematical metrics of algorithmic fairness are inherently rooted in a locally distributive conception of justice in that they are concerned with how the algorithm would allocate the relevant computational or material goods across different groups or individuals (I discuss some exceptions in the next section).[2] Computer scientists do not always and explicitly provide a moral or philosophical justification of their proposed mathematical formulation of algorithmic fairness. However, across the board, whether they are explicitly defended as such or not, most mathematical criteria for algorithmic fairness can be understood as mathematical translations of some ideals or principles for a local distribution of computational or material benefits or burdens among people.

For instance, recall the statistical criterion of calibration within groups. The purpose of this criterion is to satisfy the distribution of risk scores across different groups according to the following definition: For each possible risk score, the percentage of individuals assigned that risk score who are actually positive (the individuals who go on to commit crimes) is the same for each relevant group and is equal to that risk score. This distribution is local in the local impact of the algorithm is the locus of justice assessment. In a similar vein, the equal distribution of predictive accuracy of false positive or false negative error rates across the partition by race is a locally-relevant criterion of algorithmic justice as proposed by Angwin et al. [1].

Perhaps the most influential paradigm of distributive justice in the contemporary literature is due to John Rawls. Rawls' philosophical work has motivated an extensive program of research on algorithmic fairness. Indeed, Rawls' principles of distributive justice are so foundational to the field that he has been named 'AI's favourite philosopher' [44]. Several proposed formalizations of Rawlsian theories of fair equality of opportunity are examined by Hardt et al. [25], Joseph et al. [31], and Heidari et al. [27, 28] (among others), while portions of Rawls' theory of distributive justice are given mathematical expression in the work of Joseph et al. [30] and Hoshimoto et al. [26]. More recently, Rawls' principles of distributive justice are used to offer a theory of justice for artificial intelligence [21].

Several political philosophers have contributed to defending a version of the distributive justice paradigm by proposing an alternative theory of distributive justice or emending Rawls' original formulation [16, 17, 42]. The relevance of their works to research on algorithmic fairness is acknowledged in the algorithmic fairness literature, most notably by Binns [5]. While a thorough engagement with the substantial literature on the nuances of distributive justice is beyond the scope of this paper, it suffices to say that most fairness constraints and metrics are deeply rooted in the paradigm of distributive justice, broadly construed. Except for a few proposals for taking procedural fairness seriously [24, 29, 56], algorithmic fairness is firmly embedded in the distributive paradigm.

But would a fair distribution of an algorithm's predictions or decisions mitigate injustices raised by the algorithmic systems? Not always, if we truly pay attention to the sociotechnical nature of algorithms. In the next section, I defend this contention by arguing that structural injustices are relevant to our conceptions of algorithmic fairness. The importance of the distributive fairness metrics (locally understood) depends on the power structures and social dynamics in and through which the algorithm operates.

3 ALGORITHMIC FAIRNESS AND STRUCTURAL INJUSTICE

Consider Mandy, a low-paid employee and single mother of two underage children. Mandy lives in a fast-growing city. As the evidence attests, a notorious Tech Giant has a vested interest in automating services in various institutions in Mandy's city. The presence of the Tech Giant contributes to driving up housing costs locally, making it unaffordable for many to live in their present homes. Greater numbers of people cannot afford rent, let alone home-ownership. Driven out of their homes, the indigent workers resort to makeshift homeless shelters under road bypasses or camp out in the parks at night, just as the Tech Giant's employees, who, bolstered by generous wages, race to buy every piece of real estate in sight.

One day, Mandy's landlord lands her with a notice of eviction, intending to renovate the premises with a view to raising the rent later. It does not take long for Mandy to realize that all landlords are pursuing similarly amoral profit-seeking rents, and she is effectively priced out of the rental market. Given her limited options, she resorts to applying for a bank loan, now allocated on the basis of algorithmic services thanks to the involvement of the Tech Giant.

Apparently, the Tech Giant has made living conditions significantly worse for Mandy and most other ordinary residents, apart from the property-owning class, whose windfall gains in real estate are astronomical. Eventually, the Tech Giant has to reckon with public pressure to mitigate this starkly unjust situation (perhaps a local grassroots movement to secure affordable housing has sprung up in Mandy's town). But how?

[2]Broadly put, distributive justice is concerned with the institutional distribution of benefits and burdens across members of society [18, 36, 45]. In research on algorithmic fairness, theories or principles of social justice are often translated into the distribution of material (such as employment opportunities) or computational (such as predictive performance) goods across the different social groups or individuals known to be affected by algorithmic outputs.

One solution is to distribute a number of low-interest loans to residents. A panel of in-house hired ethicists propose that the loan allocation algorithm should comply with an equality of opportunity fairness metric. Having applied for a loan, Mandy receives the maximum amount, thanks to the distributive fairness metric embedded in the design of the decision-making algorithm. The Tech Giant's press release boasts that it complies with the ethical principles of social justice, having thus operationalized social justice into the automated decision making: "Through our firm commitment to ethical principles, we have operationalized social justice in automated decision making."

Let us examine this scenario more closely. First, the example clearly illustrates that algorithms are not purely mathematico-computational, do not operate in a vacuum, and cannot be analyzed in isolation. As several scholars have argued, algorithms are inherently sociotechnical entities (see, for example, Suchman [47], Collins [11], and Selbst et al. [46]); roughly put, algorithms are neither constructed nor adequately represented independently of the specific social contexts, institutions, power relations, and values in which they operate. On the contrary, an operating algorithm is an entity continuous with its social setting: i.e., a sociotechnical entity. This means that the data-driven predictive algorithm is not only developed within an ecosystem of socio-political and ethical values, but it is also a proper component of the institutions in which it operates. We need to take algorithmic justice in general to be concerned with such sociotechnical systems [14, 33]. Thus, fairness of algorithms is inherently (and therefore inescapably) connected with its reception in its social context of utilization and its response to the social injustice at work.

But what can a purely local resource allocation achieve against the background of power relations and social dynamics in which the algorithm operates in? Even if, by some mathematical measure, the loan allocation algorithm is deemed 'fair', it seems that the locally just decision to allocate the loan to Mandy might be fully irrelevant: it has no useful broader social significance, so long as the underlying injustice goes unaddressed. Under the conditions in which Mandy is living, it seems likely that she will accumulate a mountain of debt, instead of quickly paying off her bank loan. Overall, the initial putative 'fair' allocation will have worsened her position in the long run, despite marginally improving it for a brief time. Even if the algorithmic fairness criterion positively discriminates in favor of people like Mandy, her problem persists.

Therefore, I argue that we cannot ascertain whether 'fair' distributions computed by 'fair' algorithms have genuine moral force and significance except in comparison to an overall assessment of the kinds of injustices algorithms as sociotechnical systems are an important part of. In Mandy's case, the algorithm is embedded in the power relations between the Tech Giant and other individual and collective elements in that society and any assessment of fairness must be considered in light of the algorithms' capacity to address the sources, as well as outcomes, of injustice.

Algorithms understood in terms of their sociotechnical character can act both as the exacerbating cause and the mitigating remedy of the relevant sources and patterns of social injustice. However, not all problems of social injustice have local distributive solutions. While local distributive matters ought to be taken seriously by algorithmic fairness research, the ethical credibility of the algorithm is moot if the algorithm itself does not address the sources of structural injustice, as opposed to merely addressing its effects. That is, genuine ethical force arises from tackling social injustice at the root, not just in softening over its effects via purely and locally distributive (viz. outcome-based) solutions. Since distributive solutions are outcome-based, they are inherently incapable of preventing the perpetuation of long-lasting cycles of inequality. In sum, taking the sociotechnical character of algorithms seriously goes some way towards resolving the issues arising from the application of fairness metrics narrowly construed. That algorithms are sociotechnical processes entails that algorithms must be responsive to the relevant sources of social injustice. The broader inference, moreover, is that algorithmic interventions (however construed) must have a correct justification and motivation in order to be adequate candidates for fairness or justice.

The prominent political and feminist theorist and philosopher, Iris Marion Young, popularized the contemporary concept of structural injustice, in particular as a critique of the scope of the distributive justice paradigm [50–53, 55].[3] Young's argument begins with the observation that contemporary liberal political philosophy has been primarily and narrowly focused on distributive justice. But, she argues, beyond concerns about local distribution, unjust and wrongful harms against some people but benefits for others are also the result of structures and processes. According to Young, we might not be able to resolve these structural harms simply by altering the local distribution of burdens and benefits across the members of a society. In these cases, the locally distributive approach alone might become only cosmetic. Rather, cases of structural injustice demand that we overcome, revise, or reform the problematic structures.

Furthermore, in Mandy's type of scenario, the operative injustice is not necessarily identifiable by isolating individual responsible agents or particular algorithms and to blame them. Rather, the injustice arises as the sum of (arguably) non-blameworthy actions of multiple agents which can only be changed through structural change and collective action. Let us call the such injustices 'structural', as Iris Marion Young proposes [53]. What are these structures and when do structural injustices exist?

The structure, according to Young [53, p.111], is defined as follows:

> As I understand the concept, the confluence of institutional rules and interactive routines, mobilization of resources, as well as physical structures such as buildings and roads. These constitute the historical givens in relation to which individuals act, and which are relatively stable over time. Social structures serve as background conditions for individual actions by presenting actors with options; they provide "channels" that both enable action and constrain it.

Moreover, in a widely cited passage, Young [54, p. 52] defines structural injustice as follows:

> Structural injustice, then, exists when social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop

[3] Young's conceptions of structure and structural injustice are rooted in the works of philosophers and sociologists such as Peter Blau [6], Pierre Bourdieu [7], and Anthony Giddens [22]. See Young [53] for more details.

> and exercise their capacities, at the same time that these processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them. [...] Structural injustice occurs as a consequence of many individuals and institutions acting to pursue their particular goals and interests, for the most part within the limits of accepted rules and norms.

That is to say, structural injustice results from the operation of multiple social processes; in Mandy's case, this includes the unregulated expansion of the economic and social power of a tech giant coupled with laissez-faire policies and unregulated profit-oriented capitalism. The aggregate of factors is ultimately responsible for Mandy's situation. The situation is familiar enough, as it places individuals and groups under the systematic threat of domination or deprivation, while enabling others to dominate and exploit them.

To reiterate, Mandy's scenario illustrates an example of structural injustice caused by a supremely powerful individual company as well as other actors in positions of power (the mayor, the city council, etc.). Note that structural injustice renders individuals (and certain social groups) vulnerable to domination or oppression, even if there is no single individual to identify and blame. According to Young [55], structural injustice is a moral wrong distinct from the wrongful action of an individual agent or the repressive policies of a state. In addition, structural injustice contrasts with at least two other forms of harms or wrongs, namely that which comes about through an individual action, and that which is attributable to specific actions and policies of states or other powerful institutions. Thus, structural injustice occurs as an aggregate consequence of individuals and institutions acting to pursue particular goals and interests, largely within the limits of accepted rules and norms.

In Mandy's case, conceptualizing problems of justice as 'goods' that must be locally distributed (i.e., the fair allocation of loan) partially obscures the fact that many actions which are the roots and the reasons for the persistence of the relevant injustice result from dynamic processes, relations, and doers.

If Tech Giant is actively contributing to the unjust social conditions besetting Mandy and others like her, the algorithmic distributive fairness of the loan allocation itself is largely beside the point, since in this case it is mainly the source of the injustice (and its perpetuation) that matters. To put it differently, although some fairness criteria, e.g., those proposed to address statistical error distributions, can be justified and used locally in terms of ensuring fairer outcomes, the social significance of the fair algorithm is negligible inasmuch as it fails to address (or even recognize) the main source of the injustice in the Tech Giant's activities. To guarantee significance, the dynamic and compounding elements of injustice must be included into the assessment of algorithmic fairness. In Mandy's case, the key cause of social injustice is the negative change in living conditions brought about by the Tech Giant, and it seems that the fair algorithm – as sketched above – will not be able to amend it. The example makes clear that algorithmic fairness should not be evaluated in isolation from originating and perpetuating factors, but in connection with the structure of injustice.

To be sure, the limits of distributive fairness metrics have not been entirely neglected by the literature on algorithmic fairness. Grgić-Hlača et al. [24] proposes ways to include procedural fairness, i.e., the fairness of the decision making, into the research on algorithmic fairness. Hoffmann [29] draws heavily on legal studies in discussing the limits of the goods included in the prevailing paradigm of distributive justice in the literature. Zimmermann and Lee-Stronach [56] draw attention to the relevance of procedural justice to algorithmic systems and argue that reliance on algorithmic systems is procedurally unjust in contexts involving background conditions of structural injustice. The arguments presented in this paper are novel in that Grgić-Hlača et al. [24] does not engage with the literature on structural injustice. Moreover, while Hoffmann [29] and Zimmermann and Lee-Stronach [56] briefly quote Young's work on structural injustice, they do not explicitly engage with her account for correcting the limitations of fairness metrics. In this paper, I complement these earlier works by arguing why and how structural injustice is to be conceptually included in algorithmic fairness research.

4 TOWARDS RESPONSIBLE ALGORITHMIC FAIRNESS

Granting the broad lines of argument defended so far, the following question arises: How should sociotechnical algorithmic systems deal with structural injustice? How should the designers of the algorithm, e.g., the Tech Giant employees, and the Tech Giant itself act in light of the existence of structural injustice?

In correcting the harms of structural injustices, Young suggests that the key is to develop an appropriate account of responsibility, which she calls a *social connection model* [53, 54]. The social connection model of responsibility is forward-looking; it says what can be done when we go forward.[4]

According to Young, agents who participate in a structure of concerted action (e.g., a cooperative structure) are responsible for them in the sense that they are part of the process that results in structural injustice further down the line. In the case of Mandy, the responsible actors include the owners and designers of the algorithms, the Tech Giant as institution, its employees, and similar agents. While these actors are not necessarily responsible in the sense of having directed the process or intended the outcome, they nevertheless bear responsibility for the resulting structural injustice, because they contribute by their actions to the processes that produce unjust outcomes.

Young suggests four modes of reasoning for engaging with the social connection model of responsibility for structural injustice. According to her, responsibility for justice encourages us to think about how we can best take responsibility for reducing injustice by reflecting on these four parameters: different options of power, privilege, interest, and collective ability.

First, take the concept of power. Different levels of causal influence and capacities are relevant and contribute to changing processes. For example, a few key players may have both a greater

[4] According to Young, it is undesirable to rely on a purely legalistic liability model to restrict the range to cases in which an agent causing the harm may be identified. Young suggests to shift from mere attribution of blame and guilt to an action model. How these two models relate to each other is itself a subject of ongoing discussion. See, for example, [23].

capacity to make changes themselves and to influence the process. A few select actors are responsible for the choice of fairness metric; part of their responsibility ought to be to assess whether the selected metric removes, in any meaningful way, the operative structural injustice. If not, does the fairness principle in the context of the problem of structural injustice result in ethics washing? It is the responsibility of corporations with both vast influence and capacity to mitigate structural injustices arising operationally, in this scenario, from the expansion of machine learning algorithms or other technical systems.

Second, consider the concept of privilege. Privilege also influences one's degree of responsibility: the more privileged have a greater responsibility to address injustice. By virtue of their social position, the privileged may, like the powerful, exert an influence on merely the adoption of fairness metrics in algorithmic decision making. They hold an additional responsibility to devote some of the resources—material or otherwise—towards the instantiation of the relevant kind justice in the field.

Third, take the notion of an interest vested in the structure. Those who have a vital interest in changing oppressive structures have direct responsibilities in the order of mitigation. To the extent that the powerful have a vested interest in addressing injustice, they too must examine the broader justice consequences of algorithmic systems.

Fourth, consider collective ability. In some cases, collective organization capacities and resources exist to some extent. It is a reasonable practical policy to draw upon these (if they exist at all). In some instances, student associations, faith-based organizations, labor unions, or stockholder groups already exercise some degree of power in being able to coordinate members to take certain actions. Young proposes to harness organizational resources where doing so would prove effective.

The degrees and kind of forward-looking responsibility for changing structural processes to reduce injustice vary according to the four parameters sketched above. On Young's account, No one who participates in processes that produce structural injustice is exempt from responsibility to join with others to change those structures. Some are less inclined to do so, however, because their positions give them more interest in preserving than in changing them. Others stand in positions of relative weakness in the structures. The former must usually be pressured to take steps aimed at changing the effects of their actions, and the political responsibility of the latter often amounts to little else than organizing to criticize and pressure more powerful actors [55].

As Young continues [55], it requires organization, the will to cooperate on the part of many diverse actors, significant knowledge of how the actions of individuals and the rules and purposes of institutions conspire to produce injustice, and the ability to foresee the likely consequences of proposed remedies. One or more of these conditions is often absent. But something else often stands in the way of trying to bring about these conditions, namely the attempt by participants in the process to deny that they have a responsibility to try to remedy injustice. Future work is required to produce more awareness about such strategies of avoidance in talking to one another about responsibility and collective action, and in holding one another accountable.[5]

Given the power and privilege of many institutions who design artificial intelligence systems, it is necessary to address the four parameters of the problem as proposed by Young. That is to say, the designers and social institutions owe a four-way responsibility to address structural injustice. As illustrated and argued above, any adequate form of algorithmic justice requires responsibility in this comprehensive sense.

5 PRACTICAL CONSEQUENCES

Next, I examine the practical consequences of the arguments above for research on algorithmic fairness.

First, the motivations, justifications, and significance of operationalizing relevant locally distributional and structural elements of social injustice into algorithmic ecosystems are to be evaluated more carefully and systematically. Algorithms are to be understood as sociotechnical entities in a rich sense of the term. The range of issues that can legitimately be discussed and evaluated in relation to the sociotechnical character of algorithmic fairness must be comprehensively expanded, while finessing the representation of the relevant concepts of social (in)justice. Specifically, algorithmic fairness is to be evaluated in a broader context of algorithmic justice, encompassing both the distributive metrics and structural elements of injustices. This in part requires broadening the understanding of justice for sociotechnical systems by regarding algorithms not as mere distributors of computational or material goods, but as participants, doers, and actors embedded in complex social relations and dynamics.

Second, algorithmic outputs by informing or automating various decisions affect lives and have significant social impacts. Hence, they are intrinsically bound up with structures of social power. Discussions of algorithmic fairness tend to often overlook this power aspect of the operative context. I, however, showed that social justice concerns and formal notions of algorithmic fairness must be jointly analyzed, and that the promises and limits of the formal properties of fair predictions or decisions must be robustly related to the context of social power. A socially situated conception of algorithmic justice and fairness allows us to better trace the interdependencies between power, algorithmic reason, algorithmic rationality, and authority, and to consider how power relations affect the mode as well as the outcome of algorithmic applications.

Third, the areas of intervening to secure social justice should not be merely local and technical, at the level of data preparation, model learning or post-processing. To know whether an algorithmic output is fair it should be evaluated in trade-off with other interconnected solutions to social justice. This discussion of how to approach this trade-off problem should find central stage and

[5] Yet, unlike retrospective responsibility, forward-looking responsibility is not a matter of having caused an existing (morally problematic) state of affairs. Instead, Young's account renders responsibility a matter of being morally charged with—that is, responsible for—bringing about a state of affairs that we collectively regard as an improvement. Forward-looking responsibility is not completely removed from considerations of blame – we sometimes appropriately blame those who fail to take their responsibilities seriously—but blame is not morally salient. Rather, prospective responsibility is morally salient because we think that such responsibility may, if taken seriously by those who are held responsible, help to bring about an improved worldly state of affairs.

be taken seriously while specifying agendas for algorithmic fairness research program. To this end, future interdisciplinary work is needed to give more precision to the notion of structural injustice and to find a meaningful trade-off between its demands in the context of algorithmic predictions or decisions.

Fourth, algorithmic fairness must be treated in relation to responsibility. How should we design fair algorithms responsibly? To this end, I took some steps forward in applying Young's fourfold account of forward-looking responsibility. The interrelation between the topics of moral and political responsibility and algorithmic fairness which have gone largely neglected in the literature are to be articulated together. Responsible algorithmic fairness renders the notions of structural and distributive justice as complements to each other, not competitors.

Fifth, the proposed sense of responsible fairness brings the discussion to the notion of the responsible citizen or the responsible corporations and companies. As individual agents, overspecialization threatens to obscure the true extent of personal responsibility. A computer scientist's role, for example, is not only to design a fair algorithm by measuring some local aspects of fairness via mathematical and computational tools. In her role as a citizen, she is also responsible for working against structural injustices in her capacity. This conception requires collaboration between different ethical, social, and political groups. The broader understanding of justice brings considerations of power to the fore. These, rather than the purely formal properties of metrics, are the fundamental issues of relevance to algorithmic fairness to address.

Finally, algorithms are powerful tools with a range of uses. In application, they promote good or bad outcomes, depending upon implementation and context. I think there is the potential that algorithmic systems can be used to repair some problematic structures and to generate better ones. It is essentially the perspective of structural justice that will allow us to determine the applications of algorithmic systems with genuine moral force. It remains a question for future research to find how to fruitfully unleash this aspect of algorithmic systems.

6 CONCLUSION

Machine learning algorithms are widely used to guide the distribution of goods. Within the arena of machine learning, algorithmic fairness is a burgeoning field of research that focuses on the design of mathematical metrics for redressing harms caused by biased distributional effects of societal algorithmic systems. These metrics are often motivated by the twin political ideals of social justice and fairness. Against this background, the current paper drew on feminist political theory, and in particular the works of Iris Marion Young, in arguing that a primarily distributive approach to the local allocation of material and computational goods is too narrow to adequately address or accommodate concerns about social injustice vis-à-vis algorithmic fairness. Algorithmic fairness is, therefore, morally salient and in need of philosophical attention.

The paper argued that algorithmic ecosystems are sociotechnical entities, and so must be receptive to different sources of social injustice. The metrics of algorithmic fairness are primarily concerned with local matters of distributional justice. However, not all sources of social injustice are distributional; some are structural. After taking some steps in unpacking what responsible algorithmic fairness is, I argued for six positive corollaries of its adoption as the conceptual basis for research into the infrastructural fairness of algorithmic ecosystems and their direct effects.

This paper aimed to connect some dimensions of the philosophical works of Iris Marion Young to algorithmic fairness. I used a Youngian conception of forward-looking responsibility to sketch a way to improve the limits of algorithmic fairness. Other accounts of forward-looking responsibility have been also discussed and developed in the literature; see for example, Miller [40] French and Wettstein [20]. One next step is to make a more comprehensive engagement with the literature on forward-looking responsibility in the context of algorithmic justice. In this paper, I examined a particular approach for bridging feminist philosophy to the literature on algorithmic fairness. A complimentary exploration of the works of other feminist philosophers, from different epistemic and cultural traditions, for improving the current conceptions of algorithmic fairness is an important task to explore elsewhere.

ACKNOWLEDGEMENTS

I am grateful to Philip Pettit, Mario Günther, Iason Gabriel, two anonymous reviewers, audiences of the Philosophy Discussion Club at Cornell University, participants in the seminar on Human Rights and Global Politics at Scuola Superiore Sant'Anna Pisa, and audiences at the workshop on Responsibility & AI at the University of Vienna for valuable discussions and suggestions.

REFERENCES

[1] Julia Angwin, Larson Jeff, Mattu Surya, and Kirchner Lauren. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals and It's Biased Against Blacks.
[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.
[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.
[4] Elettra Bietti. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 210–219.
[5] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.
[6] Peter Michael Blau. 1977. *Inequality and heterogeneity: A primitive theory of social structure*. Vol. 7. Free Press New York.
[7] Pierre Bourdieu. 1984. *Distinction: A social critique of the judgment of taste*. Harvard University Press.
[8] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.
[9] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[10] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.
[11] Harry M Collins. 1990. *Artificial experts: Social knowledge and intelligent machines*. MIT press.
[12] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
[13] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Technical report, Northpole* 7, 4 (2016).
[14] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. 2021. A sociotechnical view of algorithmic fairness. *Information Systems Journal* (2021).
[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[16] Ronald Dworkin. 1981. What is Equality? Part1: Equality of Welfare. *Philosophy and Public Affairs* 10 (1981), 185–246.
[17] Ronald Dworkin. 1981. What is Equality? Part2: Equality of Resources. *Philosophy and Public Affairs* 10 (1981), 283–345.
[18] Samuel Fleischacker. 2009. *A short history of distributive justice.* Harvard University Press.
[19] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation* 80 (2016), 38.
[20] Peter A French and Howard K Wettstein. 2014. *Forward-looking collective responsibility.* Wiley Periodicals.
[21] Iason Gabriel. 2022. Toward a Theory of Justice for Artificial Intelligence. *Daedalus* 151, 2 (2022), 218–231.
[22] Anthony Giddens. 1979. *Central problems in social theory: Action, structure, and contradiction in social analysis.* MacMillan Press Ltd.
[23] Robert E Goodin and Christian Barry. 2021. Responsibility for structural injustice: A third thought. *Politics, Philosophy & Economics* (2021), 1470594X211027257.
[24] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[25] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).
[26] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.
[27] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*. 1265–1276.
[28] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 181–190.
[29] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.
[30] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2016. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559* 1, 2 (2016), 19.
[31] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems* 29 (2016).
[32] Atoosa Kasirzadeh and Damian Clifford. 2021. Fairness and Data Protection Impact Assessments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 146–153.
[33] Atoosa Kasirzadeh and Colin Klein. 2021. The Ethical Gravity Thesis: Marrian Levels and the Persistence of Bias in Automated Decision-making Systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 618–626.
[34] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801.
[35] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).
[36] Julian Lamont. 2017. *Distributive justice.* Routledge.
[37] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9, 1 (2016), 3–3.
[38] Martinez, Emmanuel and Lauren Kirchner. 2021. The Secret Bias Hidden in Mortgage-Approval Algorithms. https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms, Last accessed on 2022-02-24.
[39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
[40] David Miller. 2008. National responsibility and global justice. *Critical review of international social and political philosophy* 11, 4 (2008), 383–399.
[41] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.
[42] Robert Nozick. 1973. Distributive justice. *Philosophy & Public Affairs* (1973), 45–126.
[43] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.
[44] Ariel Procaccia. 2019. AI Researchers Are Pushing Bias Out of Algorithms. https://www.bloomberg.com/opinion/articles/2019-03-07/ai-researchers-are-pushing-bias-out-of-algorithms, Last accessed on 2022-02-24.
[45] John Rawls. 1971. *A theory of justice.* Oxford: Oxford University Press.
[46] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.
[47] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication.* Cambridge university press.
[48] Ben Wagner. 2018. Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? In *Being Profiled*. Amsterdam University Press, 84–89.
[49] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks Posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088
[50] Iris Marion Young. 2000. *Inclusion and democracy.* Oxford University Press.
[51] Iris Marion Young. 2001. Equality of whom? Social groups and judgments of injustice. *Journal of political philosophy* 9, 1 (2001), 1–18.
[52] Iris Marion Young. 2002. Lived body vs gender: Reflections on social structure and subjectivity. *Ratio* 15, 4 (2002), 410–428.
[53] Iris Marion Young. 2006. Responsibility and global justice: A social connection model. *Social philosophy and policy* 23, 1 (2006), 102–130.
[54] Iris Marion Young. 2011. *Justice and the Politics of Difference.* Princeton University Press.
[55] Iris Marion Young. 2011. *Responsibility for justice.* Oxford University Press.
[56] Annette Zimmermann and Chad Lee-Stronach. 2021. Proceed with Caution. *Canadian Journal of Philosophy* (2021), 1–20.