Received: 13 November 2020 | Revised: 23 May 2021 | Accepted: 31 May 2021

DOI: 10.1111/phc3.12760

ARTICLE

WILEY

Algorithmic bias: Senses, sources, solutions

Sina Fazelpour[1] | David Danks[2]

[1]Department of Philosophy and Religion, Northeastern University, Boston, Massachusetts, USA

[2]Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence
Sina Fazelpour, Department of Philosophy and Religion and the Khoury College of Computer Sciences, 360 Huntington Ave, Northeastern University, Boston, MA 02115, USA.
Email: s.fazel-pour@northeastern.edu

Funding information
Social Sciences and Humanities Research Council of Canada, Grant/Award Number: #756-2019-0289 2

Abstract

Data-driven algorithms are widely used to make or assist decisions in sensitive domains, including healthcare, social services, education, hiring, and criminal justice. In various cases, such algorithms have preserved or even exacerbated biases against vulnerable communities, sparking a vibrant field of research focused on so-called algorithmic biases. This research includes work on identification, diagnosis, and response to biases in algorithm-based decision-making. This paper aims to facilitate the application of philosophical analysis to these contested issues by providing an overview of three key topics: What is algorithmic bias? Why and how can it occur? What can and should be done about it? Throughout, we highlight connections—both actual and potential—with philosophical ideas and concerns.

1 | INTRODUCTION

Key social and personal decisions that impact our lives are increasingly guided by predictive algorithms. Medical diagnoses incorporate predictive models built using large datasets; loan approvals are informed by algorithmic judgments of credit worthiness; decisions to send social workers to investigate potential child abuse are guided by algorithm-based risk scores; and the examples multiply every day. At the same time, there is increasing awareness of the harmful impacts caused by biases in these algorithms: face recognition algorithms that perform worse for people with feminine features or darker skin (and worse still for those with both) prevent people from accessing resources; recidivism prediction models that rate people of color as significantly higher risk than white counterparts lead to unjust continued imprisonment; and many more. In each case, the use of algorithms threatens to preserve, or even compound, existing injustices.

[Correction added on 03 December 2021, after first online publication: The copyright line was changed.]

Philosophers have begun to help characterize, diagnose, and mitigate these algorithmic biases through both critical conceptual work, and also applied efforts to respond to the impacts of algorithmic bias. In this paper, we consider three key questions about algorithmic bias: What is it? How does it arise? What can or should be done about it? While exploring each question, we indicate places where there is current philosophical research, and also where more philosophical insight would prove helpful. Algorithms and their biases have already changed many of our lives, and their importance will only increase in coming years.

Philosophical engagement with algorithms in general, and the challenges posed by them, extends to many topics that are beyond our focus on algorithmic bias. Of course, there is not always a bright line that can be drawn to distinguish philosophical work on algorithmic bias from those other topics, such as how non-epistemic values might influence model selection in machine learning (Dotan, 2020). At the same time, space constraints necessarily require the omission of many important and related topics including trust (Danks, 2019), privacy (Véliz, 2020), transparency (Creel, 2020), and much more. Algorithmic bias is only one issue within the broad collection of philosophical topics often lumped together under the heading of 'Ethics and AI'. Our focus will be algorithmic bias, but we do not thereby mean to suggest that these other issues are unimportant or uninteresting, whether philosophically or societally.

Throughout this paper, we use a running example to illustrate the complexities. Many universities increasingly use data from past students to build models for predicting 'student success' (defined in various ways), where those models can support informed changes in policies and practices. This situation is a textbook example of modern machine learning: collect data on past cases (e.g., students); train a model on these data that optimizes predictions about an outcome of interest (e.g., probability of success); and use that predictive model to guide action (e.g., policies for mentorship or admission strategies). These kinds of algorithms are being built at universities worldwide, likely including those with which the reader has some past or present affiliation. Of course, universities might have multiple goals for these models, including helping students, reducing attrition costs, improving the university's reputation, increasing government funding, or others (Delen, 2010). As we discuss below, those goals can impact everything from how we measure, to what we optimize, to which actions we pursue, to which people are benefited or harmed.

Finally, a terminological note: data scientists typically apply a learning algorithm *L* to a dataset *D* to build a predictive model *M* that can function as a predictive algorithm. In practice, discussions about algorithmic bias are essentially always about *M*, the predictive algorithm/model. As such, we will use the terms 'algorithm' and 'model' interchangeably in this paper to talk about *M*. When we are concerned specifically with *L*, we will explicitly talk about the 'learning algorithm.'

2 | SENSES OF BIAS

We begin with the question: What is algorithmic bias? In everyday usage, the term 'bias' typically carries negative connotations, but it is more ambiguous when talking about *algorithmic* bias. At its most neutral, algorithmic bias is simply systematic deviation in algorithm output, performance, or impact, relative to some norm or standard (Antony, 2016; Danks & London, 2017; Johnson, 2020). An algorithm can be morally, statistically, or socially biased (or other), depending on the normative standard used. For example, our model's predictions of student success could be statistically biased if those predictions differ from previously observed training data; predictions that systematically deviate from past history encode a measure of statistical bias. Alternately, the predictions could be morally biased; for example, they might depend in illegitimate ways on the gender of the student.[1] More generally, specificity about the normative standard enables more nuanced evaluations: not all statistically (or legally) biased behaviors are ethically or morally problematic, while not all statistically fair or unbiased predictions are ethically or morally acceptable. Moreover, algorithmic bias is not a function solely of the code or mathematics, but also depends on the domains of application, goals for the algorithm use, and other contextual factors. The same student support algorithm (in the sense of underlying code or mathematics) could be unbiased at one university, but biased—statistically, morally, socially—if used at a different university. In this paper, we primarily use the term 'algorithmic bias' in the more neutral sense, but our examples and focus will typically be on the more salient morally problematic instances.

One might object that algorithms are only mathematical objects, and so must be completely objective. This *neutrality thesis* holds that technological artifacts such as algorithms do not embody or implement values, and so their *use* is the only appropriate locus of ethical evaluation. By analogy, one would typically not say that a hammer is biased, though it can be used in biased ways. Much work in philosophy of technology and value-sensitive design has cast significant doubt on the neutrality thesis in general (e.g., Dotan, 2020; Friedman & Hendry, 2019; Simon, 2017). In particular, algorithms implement values, and so potentially biases, in at least two high-level ways. First, algorithms make possible certain kinds of decisions or capabilities, and so embody the value that those decisions or capabilities are important. The decision to employ data-driven algorithms in a given domain can itself have social, political, and economic significance, as the very existence of the algorithm can disproportionately and unjustly favor one group over another (Liao & Huebner, in press; see also Winner, 1980). The widespread use of algorithms across many aspects of society can act to rigidify unjust and oppressive social practices and structures, regardless of the exact content of the predictive model (Castro, 2019; O'Neil, 2016).

Second, algorithms implement values because they are almost always optimized for performance relative to a standard. They are developed to succeed according to some metric (e.g., predictive accuracy, sensitivity, specificity, F1-score), and the algorithm thus values or promotes whatever counts as 'success' on that metric. Algorithms are not objective, but rather embody the value-laden view that some performance is better or more important than others. The algorithm developer's selection of a particular performance metric is thus key in determining these implemented values. This issue clearly connects with debates in the philosophy of science about the value-ladenness of science (Douglas, 2009). In each case, the products of inquiry—algorithms or scientific theories—are sometimes thought to be value-free, at least in their ideal forms. But in each case, their creation requires many choices that are informed by the inquirer's values, particularly about tradeoffs (e.g., accuracy vs. simplicity) or evaluation standards. Of course, algorithms and scientific practice are not thereby a matter of arbitrary choice. Rather, different algorithms, models, or theories might be appropriate depending on the particular context, desiderata, and (value) constraints. Hence, data scientists, machine learners, and others building these decision support systems should be explicit about the values (e.g., optimization criteria) that are implemented in their algorithms.

The implementation of values in algorithms also connects with debates in ethics outside of algorithmic bias. First, there are general normative ethical questions about which values ought to be prioritized or implemented in our algorithms, and so which biases (if any) ought to be permissible. Second, algorithmic biases can lead to significant harms and injustices, particularly if less-important moral values are prioritized. Much of the focus here has been on comparative harms (i.e., one group is unjustly favored by the algorithm over another), but non-comparative harms (i.e., everyone is made worse off) can also result (Eubanks, 2018; Fazelpour & Lipton, 2020; O'Neil, 2016). Hence, there are many challenges in applied moral and political philosophy around diagnosis of the harms of algorithmic biases, and how (or whether) we morally ought to mitigate or eliminate them. In our running example, ethical questions about algorithmic bias concern both which values involving student success should be instantiated in, or supported by, our algorithm (first type), as well as how we ought to respond to potential harmful biases due to the algorithm (second type).

These ethical questions are particularly challenging in a value pluralistic society. As algorithms increasingly control aspects of public life, including allocation of public goods and resources, epistemic and moral peers are increasingly likely to disagree about exactly which values should be implemented or prioritized. Student success provides a clear case of such pluralism, as members of a university community presumably place different weights on overall retention, success of students who were unlikely to succeed, remedying historical inequities that hinder the success of some groups, expected return on resources, and other values. Political philosophy has long wrestled with questions about how states or organizations ought to determine policies given value pluralism and value disagreement (Muldoon, 2016; Pildes & Anderson, 1990), and these questions arise equally for algorithm design (Berendt, 2019). In fact, the challenges of value pluralism can be even worse for algorithms. Different values in political settings sometimes imply the same policy (e.g., if values A and B are each advanced by policy P, then any

relative weighting of A versus B will have the same policy implication). In contrast, different values almost always imply different objective functions for learning a model, and so almost never result in the same algorithm. Even small differences in values lead to underdetermination of algorithms (and possible biases).

3 | SOURCES OF BIAS

Given this characterization of algorithmic bias, we now ask: How and why does it arise? The full, complicated pipeline of algorithm design, development, and deployment is depicted in Figure 1. We deliberately show some details than are not usually philosophically relevant, as popular presentations of machine learning and data science sometimes mistakenly suggest that algorithm development is a relatively simple pipeline. At the same time, given that complexity, our discussion of sources of bias across this pipeline cannot be exhaustive. The many, often interdependent choice points—some seemingly innocuous—require value judgments, and so imbue the model with values. A key general point is that the ethical and epistemic evaluation of an algorithm must focus on the process, not simply the final technological product.

3.1 | Biases in problem specification

The first step in algorithm design is problem specification: What goal(s) will the algorithm be used to achieve (perhaps as part of a broader system)? This specification requires thinking about our overall aims, the actions available to us, and ways of using the algorithm to help achieve those aims (Mitchell et al., 2021). In most practical settings, decision makers are interested in aims that are complex, contested, and sometimes intentionally ambiguous. Each of these elements requires consideration of values and normative standards, and thereby provides a vehicle for the creation of biases.

Consider our running example: universities must determine what they aim to achieve, including clear specification of what counts as 'student success' in terms of *target* variables to be predicted. The translation of the complex and ambiguous outcome of interest into quantifiable surrogate target variables requires significant domain-specific expertise and value judgments (Passi & Barocas, 2019). Student success could be specified in terms of many different target variables, including grades, respect from fellow students, employer prestige, post-graduation salary, and much more. This process of translating a vague goal into precise target variables is often a matter of contestation with bargaining and negotiating between various stakeholders (Coyle & Weller, 2020), precisely because different specifications can have different impacts. In particular, there can be *disparate impacts*, where members of a protected group are differentially impacted relative to a (more) dominant group (Obermeyer et al., 2019). For example, if predicted 'success' is measured by grades during a student's first year, then enrollment or financial aid decisions based on that prediction might have a disparate *negative* impact on minority students (violating anti-discrimination legal norms; Barocas & Selbst, 2016) compared to using grades during the second year (e.g., if minority students face additional challenges in the first year; Kleinberg 2018b). Importantly, if the exact same code for the algorithm is used differently (e.g., to guide student support decisions), then we might instead have disparate *positive* impact (cf. Barabas et al., 2018).

Further harmful algorithmic biases can emerge at the system-level if our problem specifications or target variables fail to capture our complex real-world goals (Mitchell et al., 2021). For example, an algorithm that predicts which individual students are at risk of low grades would naturally support individual-level interventions that may omit key group-level goals (e.g., diversity; see Anderson, 2010). More generally, *omitted payoff bias* arises when we omit key goals (or sources of "payoff") from our target variables, which can be particularly common with complex real-world goals (Kleinberg et al., 2018a). Moving out from a specific algorithm, the very decision to employ predictive algorithms as (part of) the solution to these complex problems can alter institutional dynamics (Aizenberg &

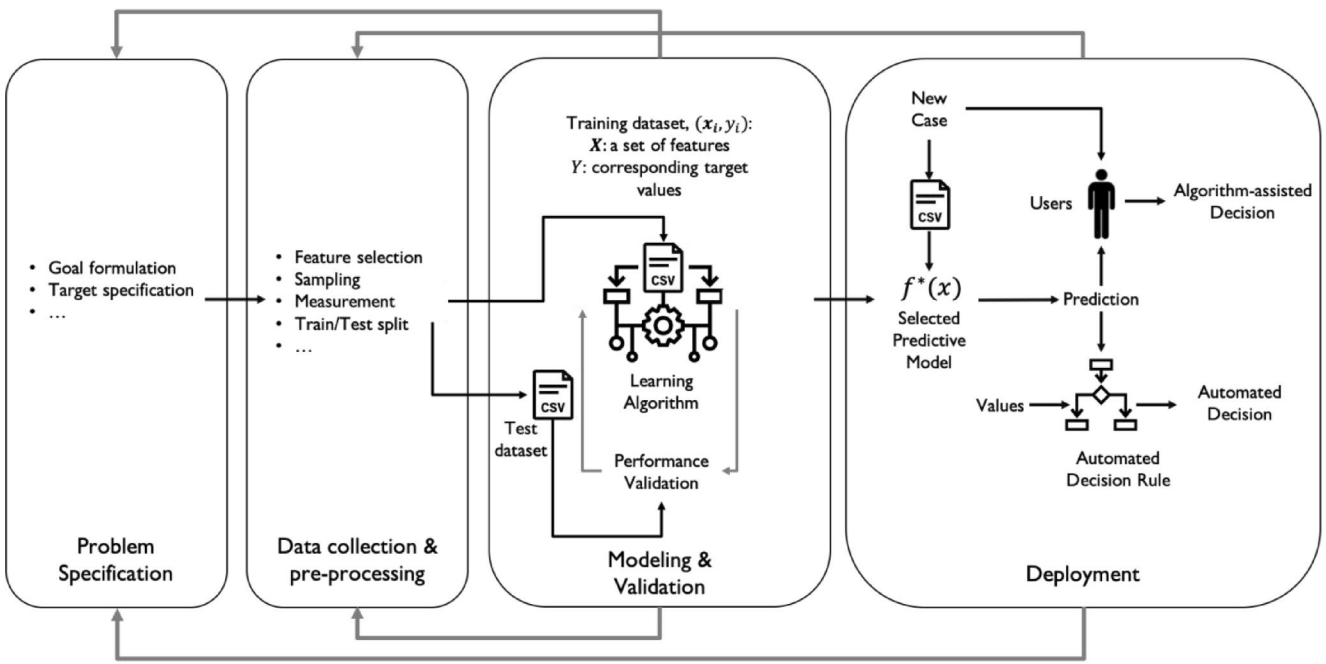

FIGURE 1 Algorithm-driven and -assisted decision-making pipeline

van den Hoven, 2020) and reshape core societal values (Merry, 2016). Philosophers should clearly inform and guide these debates, particularly about the conditions in which such changes are ethically or socially defensible.

3.2 | Biases in data

Once we determine target variables to predict or optimize, standard practice for algorithm or model development is to apply a learning algorithm to a dataset of historical cases in order to discover predictively useful patterns. Much of the technical research in machine learning aims to create increasingly powerful learning algorithms and representational frameworks for this kind of problem. Essentially, all such learning algorithms aim to produce models that reflect, at least partially, statistical features of the input dataset; that is, basically all machine learning algorithms output models that partially 'mirror' the statistics in the historical data (perhaps adjusted by our background knowledge, goal- or context-specific factors, or other non-statistical information). Hence, when there are biases (of some sort) in the historical data, then the resulting algorithm or model will typically reflect them. We can usefully separate those data biases into issues with the (measured) real-world systems, and issues with data collection procedures (Danks & London, 2017; Mitchell et al., 2021).

First, algorithmic biases can result from existing biases in the real-world systems that are measured in the data. This source of bias is the most commonly-cited one in public discussions about algorithmic bias, and is colloquially captured in the slogan "bias in, bias out" (Courtland, 2018; Mayson, 2018; cf. Rambachan & Roth, 2020): any biases in the measured world will typically be captured in the input data, and so embodied in the resulting models that 'mirror' those input data. For example, if there is systemic racism in a university that impacts student success, then our historical data will show that members of under-represented minorities are less likely to succeed, even when other causes are the same. If we use a standard learning algorithm to develop a predictive model, then that model will learn that these students have a lower probability of success. Of course, those biased predictions are not reflective of any intrinsic features of the student, but rather are due to the racist system within which past students (i.e., data subjects in our input data) have functioned. Data scientists often aim to collect and analyze representative datasets that measure how the system actually is. But if the reality of the world does not align with our normative ideals, then learning from those data can be a recipe for perpetuating harmful stereotypes and inequalities. Many notable cases of algorithmic biases involve the algorithm mirroring structural inequities and biases that are unfortunately widespread in society. In fact, there are even recent efforts to use the presence of algorithmic bias as a 'signal' of previously unnoticed real-world biases that should potentially be ameliorated (Carton et al., 2016; Pierson et al., 2020).

Second, biased data can arise from limitations and biases in our measurement methods. The simplest case is when we have non-representative input data, often leading the resulting algorithms to exhibit worse performance on the under-sampled groups. For example, if we build a dataset of faces by automatically downloading pictures of celebrities, then any resulting model (e.g., for feature identification or facial recognition) will likely underperform on members of groups that are under-represented among celebrities (Buolamwini & Gebru, 2018). In our running example, if student success is partly measured by self-reported need, and if perceived need is measured with surveys handed out in a particular location in campus (e.g., counseling services), then students who do not regularly visit that location (for whatever reason) will be much less likely to be represented in the input data (see Austin et al., 1990). The resulting predictive algorithm will likely under-perform, perhaps significantly, on those individuals; that is, the algorithm will likely be biased against them.

A different data collection challenge is that historical data rarely reveal counterfactual outcomes. For example, our student success dataset will not have information about the performance of applicants who were denied admission. This so-called *selective labels problem* (Lakkaraju et al., 2017)–the observed values of target variables in the data are a selective function of past decisions[2]–can result in morally problematic biases in our data (Kallus & Zhou, 2018). Even when the variables are observable (perhaps through various assumptions), *label bias* can arise

when our measurement procedures are themselves biased (Corbett-Davies & Goel, 2018). For example, if measurements of student success incorporate informal human assessments, then our algorithm will be optimized to predict these (potentially biased) human judgments. More generally, algorithm developers often aim to predict the so-called 'ground truth' of what is really the case. In many contexts, though, the ground truth itself lacks construct validity: it is a biased, and so imperfect, proxy for the underlying phenomenon of interest (e.g., academic performance or respect from fellow students). Importantly, this latter source of algorithmic bias is independent of real-world biases in the actual system under study. We can have label bias due to biased perceptions even if there are no underlying differences in, say, performance between two individuals (see Madera et al., 2009). Label bias can actually lead to the creation of new biases, not just the perpetuation of existing ones, if future decisions are made on the basis of the biased measurements.

3.3 | Biases in modeling & validation

Given a dataset, developers use a learning algorithm to 'fit' a predictive model on a portion of the data–the 'training set'–whose performance is validated on a previously unseen portion of the dataset–the 'test set'.[3] These modeling and validation processes, often conducted iteratively, almost always optimize and evaluate model performance relative to some criteria of success. While this stage can be highly technical, it also involves multiple value judgments. As noted earlier, the choice of objective function introduces values into our algorithms. Moreover, fine-grained details about this choice can matter: for example, we could focus on pure predictive performance (to minimize errors), or on the distribution of errors (e.g., to minimize false positives or predictions of "no help needed" for students who truly need help; see Thammasiri et al., 2014). In practice, developers often delegate these choices to the decision-makers who requested the algorithm.

A second type of value judgment arises because often many such performance metrics cannot be perfectly satisfied in tandem. These types of value judgments are familiar in the context of discussions of ethical-epistemic tradeoffs in philosophy (e.g., Gendler, 2011; Dotan, 2020; Johnson, in press). Consider Gendler's dilemma: Should we use information about, say, biased cultural practices or categories to improve our epistemic/predictive performance (but support ethically problematic practices), or ignore that information on ethical grounds (but bear an epistemic cost as a result)? Developers are generally cognizant that these value judgments cannot be resolved in a purely technical way; the choice depends on the aims and values of users of the model (Kearns & Roth, 2019).[4] Moreover, these tradeoffs are all heightened when algorithms are used repeatedly for multiple decisions, and so we may have to decide whether to allow some short-run ethical harms in order to gain additional knowledge that can enable long-term reduction in ethical harms.

A final source of potential bias paradoxically arises from efforts to reduce bias. Many proposed fairness metrics (see Section 4) provide additional, normatively justified, performance criteria for optimization in modeling and validation. Various value judgments thus arise about choice of fairness metrics. For example, the learned models might have slightly worse accuracy in order to be, in some sense, fairer (though see Dutta et al., 2020). We discuss these types of adjustments and tradeoffs below, but note here that they create new opportunities for bias.

3.4 | Biases in deployment

Finally, algorithmic biases can arise during the deployment of a predictive algorithm. As noted earlier, algorithms implement values based on what they were trained to optimize, and important biases can arise if the users' values and the algorithm's values are substantively different (Danks & London, 2017). Moreover, such failures of value alignment can happen without awareness, as the algorithm might be described to users simply as 'predicting student success' without indicating what constitutes success. For example, if users care about students who are

most likely to have low grades but the prediction algorithm is optimized to identify those likely to drop-out, then the algorithm outputs will not provide the right information. Moreover, depending on the situation, this divergence might disproportionately impact members of already disadvantaged groups.

Biases due to failures of understanding can also arise when a purely predictive (observational) model is used to make policy changes or decisions (Caruana et al., 2015). Variables or properties can predict an outcome even if they are not causes of it; the symptoms of a disease are predictively valuable, but treating those symptoms does not typically address the root cause. In the case of student success, enrolling in advanced classes might predict higher likelihood of success, but it would surely be a mistake to encourage every student in remedial courses to instead take more advanced ones in an effort to raise their probabilities of success. More generally, we need to ensure that our algorithms and models provide the correct information–whether predictive, diagnostic, or causal–for our goals and available actions, and that this information is communicated appropriately to end-users (London, 2019).

We also must ensure that the deployment context is relevantly similar to the historical context of data collection (or know how they differ). Models used in very different settings where the relevant causes are different can have quite biased outputs, both statistically and ethically. For example, a student success predictive algorithm trained on historical data from a small private university in the United States would not necessarily generalize to a large public university in China. Contextual differences become increasingly challenging (as a potential source of bias) as the algorithm is deployed for an extended period of time. Algorithms are frequently used in complex social systems with numerous feedback loops that shape data evolution (Chouldechova & Roth, 2020); for example, today's decisions about student support resources will influence next year's student population which will influence which students succeed (and which need resources). The use of the algorithm can itself change the deployment context in ways that introduce bias or other ethical complications (Benjamin, 2019; O'Neil, 2016).

In addition, a deployed algorithm might function only or primarily as decision support for humans who ultimately make decisions. Student support might be determined not directly by an algorithm, but rather by an administrator who receives algorithmic guidance. In such cases, the epistemic and ethical quality of the overall decision will depend not only on properties of the algorithm, but also on how people understand and integrate its outputs into their deliberations. However, people's uses of algorithmic outputs will differ based on decision context (Kleinberg et al., 2018a), trust (Dietvorst et al., 2015), institutional accountability frameworks (De-Arteaga et al., 2020), decision focus (Green & Chen, 2019), and more. As a result, morally problematic decisions and unjust harms could result from a biased algorithm supporting an unbiased human, an unbiased algorithm supporting a biased human, or both being biased.

Broadening our scope even further, the immediate and long-term impact of an algorithm will also be shaped by interactions with other institutions (e.g., pre-university education) or individuals (e.g., admissions personnel), including through complex feedback loops (cf. Ensign et al., 2018; Milli et al., 2019). These broader interactions can create unintended *incentive effects* when people strategically change their behaviors in response to features of the algorithm (Dai et al., 2021). For example, the student success algorithm might lead to support only for students below a specific grade average, thereby giving students an incentive to worsen their grades (so they can qualify for support). Alternately, the algorithm might value those who have already sought out help (since they might be more likely to take advantage of additional offerings), thereby disproportionately benefiting students with the knowledge, time, and/or ability to seek support, regardless of their need. More generally, these instances of Goodhart's law[5] reveal the risks of deploying algorithms without carefully considering the resulting dynamics.

4 | SOLUTIONS TO BIAS

We now turn to the prescriptive question: What can or should we do about algorithmic biases? One common response is to simply declare "Algorithms ought not use sensitive information about group membership"; one might hope that an algorithm could not possibly be biased about property *X* if it is never told about whether an

individual possesses *X*. For example, if we are concerned about racial bias in our student success prediction algorithm, then we might simply design our algorithm by removing explicit racial identifiers from the input training data.

In the US legal system, this kind of 'fairness through unawareness' is typically motivated on anti-classification grounds that aim to prevent direct discrimination or disparate treatment. Explicit use of protected attributes–legally prescribed attributes such as race, gender, and nationality–often prompts strict scrutiny by US courts (for a thoughtful critique of generic interpretations of this, see Hellman, 2020, Section 3). Unfortunately, this kind of 'fairness through unawareness' almost never succeeds. Other variables in the input data will often be correlated with, and so serve as proxies for, protected attributes; even when these sensitive attributes are not explicitly included in the data, they are 'redundantly encoded' in these proxies (Dwork et al., 2012; Johnson, 2020). For example, home zip codes are correlated with racial identification in the US. Hence, if we include zip codes in our algorithm, then they can serve as proxies for racial identifiers, thereby leading our "unaware" algorithm to nonetheless mirror historical and ongoing patterns of racism. Moreover, algorithms often put special weight on predictors for the group in numerical majority (Dwork et al., 2012; Kearns & Roth, 2019), and so "fairness through unawareness" can disadvantage members of smaller groups. And of course, explicit use of sensitive information is only one possible source of bias.[6]

We cannot eliminate algorithmic bias through unawareness. Similarly, while adopting better measurement strategies can lead to fairer algorithms (Hellman, 2020), the approach still fails to account for many other sources of bias. Recent years have thus seen an explosion of (primarily technical, but also interdisciplinary) research on 'fair machine learning (ML).' Most of this research has followed a standard strategy (discussed below), but it is important to recognize the limits of this strategy. In particular, when biases are deeply entrenched in a particular organizational or societal setting, then the proper response should often be to try to fix the underlying real-world problem (Antony, 2016; Barabas et al., 2018; Mayson, 2018), rather than turning to the technical methods of fair ML. Algorithmic bias is not a purely mathematical problem, and responses to it often must engage with the messy complexities of the real world.

4.1 | The standard debiasing strategy

Most mainstream work in fair ML pursues a two-stage strategy for addressing algorithmic bias: (1) Use one (or more) mathematical fairness measures to quantify the amount of bias in the algorithmic output; and (2) Develop mitigation responses that reduce, and ideally eliminate, bias according to that measure. Different fair ML methods result largely from variation on one of these two dimensions (see Barocas, Hardt, & Narayanan, 2019 for a detailed exposition of many of the technical features of the standard strategy).

Consider first the different proposed measures of bias. Three families of fairness measures, distinguished primarily by their mathematical language, have been prevalent in the literature: (i) individual-based, (ii) statistical, and (iii) causal and counterfactual. Individual-based fairness measures aim to mathematically quantify the Aristotelian normative ideal of treating like cases alike (Dwork et al., 2012; Dwork et al., 2020). As oft-noted by philosophers and legal scholars, the core challenge for this ethical ideal is determining what counts as "relevant similarity" with respect to the standards of a task (Hart, 1961; Westen, 1982). In practice, there are no mathematically precise, context-independent functions that appropriately quantify similarity between two individuals, so these measures are not widely used. However, given the long tradition of philosophical work on this ideal, there may be opportunities for philosophers to guide the design and selection of context-specific measures.

Statistical (or group-based) fairness measures are the most common ones, and track statistical disparities (e.g., accuracy, or error rates relating to sensitivity or specificity) between algorithmic predictions across different protected attributes that we typically think morally ought not impact decisions. For example, we might think that

student success predictions should not depend (in one of several possible ways, such as error rates across different groups) on morally arbitrary attributes such as race or gender. Many different statistical measures have been proposed, each with normative claims that it measures the statistical disparities that must be avoided in an ethically unbiased algorithm. These statistical disparities are often formalized as ratios or differences between some group-based metric; different ones are connected to concepts[7] such as: equality of treatment (e.g., Dieterich et al., 2016), the legal doctrines of disparate treatment (e.g., Berk et al., 2021) and disparate impact (e.g., Feldman et al., 2015), and the ideal of equality of opportunity (e.g., Hardt et al., 2016). These measures all use only the observable statistics of the deployment population and the algorithm outputs, rather than domain-specific understanding (except about which variables are protected in a sector or a legal system). Unfortunately, a variety of impossibility results have shown that these measures are incompatible with one another in a wide range of conditions (Chouldechova, 2017; Kleinberg et al., 2016). For example, when two groups differ in their base rate for the target variable, then it is impossible to jointly satisfy parity in group-wise predictive value and group-wise error rates.[8] The core conceptual issue is that each statistical measure prioritizes its own kind of 'equality' across the algorithm predictions (in a given context), but we can only achieve all of these equalities simultaneously if the relevant protected groups are homogenous or indistinguishable (in the relevant statistics). We thus must make a normatively grounded value judgment about which statistical measure to prioritize and minimize.

Finally, causal and counterfactual measures are motivated by the idea that questions about justice and discrimination are causal in nature, and so we must attend to the causal reasons for patterns of injustice, including what would have occurred (counterfactually) if the biases had not been present (Kusner et al., 2017; Kusner & Loftus, 2020).[9] In particular, these measures aim to eliminate algorithmic bias due to the causal influence of protected attributes on the algorithmic predictions or resulting decisions, regardless of the predictive power of those protected attributes. Causal and counterfactual measures are thus more epistemically demanding as we must know (or infer, or estimate) the relevant causal information. For example, we might need to know whether a student's gender has historically been a (morally inappropriate) *cause* of success, not just whether it is predictive of such success. Variation in causal measures arises from their granularity about causal pathways (e.g., Loftus et al., 2018 vs. Nabi & Shpitser, 2018), and their focus on population- or individual-level causal impacts (e.g., Kilbertus et al., 2017 vs. Kusner et al., 2017).

Given a fairness measure from one of these three types, one must then decide how to enforce the ethical constraints that it formalizes. One can pre-process the historical data so that it is closer to what it ethically 'should have been' (e.g., Feldman et al., 2015; Kamiran & Calders, 2012). Alternately, one can adjust the learning algorithm so that it explicitly incorporates fairness as a value, usually by including the fairness measure as an additional constraint on optimization (e.g., Zafar et al., 2017; Zhang & Bareinboim, 2018). There are different benefits to these different mitigating strategies, but both presuppose that harmful *ethical* biases can be mitigated by introducing *statistical* biases in design and development (rather than diagnosing and directly responding to the underlying causes). These strategies also often require explicit use of group membership information, as well as difficult decisions about how to potentially tradeoff ethical, epistemic, and statistical considerations.

The standard approach in fair ML requires many value choices. The impossibility results imply that we almost always must choose between fairness measures. Different ways of incorporating fairness constraints into the algorithm development pipeline can yield quite different models. Mitigation strategies can lead to significant social costs depending on the degree to which fairness gains are 'traded off' for epistemically valuable measures (e.g., predictive accuracy) that also matter for societal welfare (Corbett-Davies, Pierson, et al., 2017; Gendler, 2011; Kleinberg et al., 2016). And these are just the most obvious ways in which fair ML requires people to make choices that prioritize some values, and thereby perhaps only shift the biases in our algorithms, rather than eliminate them. Although mathematical techniques have been useful in formalizing these incompatibilities and trade-offs, choices about how to resolve them require philosophical arguments and considerations that fall outside of the narrow technical scope of the standard approach to fair ML.

4.2 | Foundational questions about responses

Philosophical research directly on fair ML can be roughly divided based on its stance towards the standard approach. Some research seeks to clarify and improve the often-implicit normative underpinnings and commitments of various fairness measures (Glymour & Herington, 2019; Hellman, 2020; Leben, 2020). Many fairness measures are purportedly based in moral and legal doctrines (e.g., 'disparate impact', 'equality of opportunity', …), and so we might naturally expect that those statistical measures would track the conceptual intuitions in a wide range of cases. In practice, however, the mathematics of a particular measure is not always precisely connected with the justifying concepts and principles, rendering the measure morally irrelevant (Hellman, 2020). Philosophical clarification of the underlying normative principles may thus provide reasons to prefer a fairness measure in a given context.

A second line of philosophical research fundamentally challenges the standard approach as an insufficient or inappropriate way to reduce algorithmic bias and injustice (Fazelpour & Lipton, 2020; Hoffman, 2019; Selbst et al., 2019). Works in this line often argue that the standard debiasing strategy incorporates unrealistic idealizations and abstractions that render it ill-suited for its intended purposes. For example, the narrow focus on static prediction in the standard approach ignores epistemically and ethically critical aspects of deployment dynamics (see Section 3.4). As a result, the standard approach can even lead to *increased* long-term and system-wide harms to the very populations it aims to protect (Dai et al., 2021; Liu et al., 2018; Hu & Chen, 2018). This line of research thus broadens the realm of normative questions and challenges that must be addressed for fair ML and the responsible use of algorithms.

Finally, one might wonder whether all of these worries about algorithmic bias provide reasons to simply reject algorithmic decision-making in the first place. We suggest that this conclusion would be much too quick. Algorithms can provide moral or societal benefits; they are not *necessarily* problematic. They can be used to identify existing biases in human practices (e.g., Carton et al., 2016), improve medical diagnoses (Haenssle et al., 2018), and so much more. Even biased algorithms can be superior to the human practices that they replace or inform (Corbett-Davies, Goel et al., 2017). Moreover, some aspects of algorithmic bias (e.g., the impossibility results for statistical measures of bias) are mathematical theorems, and so potentially apply to *any* structured decision-making, whether human- or machine-based (Tuana, 2010). The hype around algorithms and AI is frequently too loud and overwrought, and as we have seen here, algorithms can certainly create and perpetuate massive injustices. Nonetheless, we suggest that the proper response is to find ways to use algorithms to reduce inequities and injustices, not to stop using computational algorithms altogether.

5 | FUTURE DIRECTIONS

Previous sections noted where philosophers have or could fruitfully contribute to discussions of algorithmic bias, and we close by pointing to additional connections. Many of the core concerns about algorithmic bias involve concepts such as race, gender, and disability. These concepts are usually operationalized with modular and context-insensitive classification and identification schemes (e.g., Hu & Kohler-Hausmann, 2020). Work on socially-oriented metaphysics and epistemology (e.g., Beeghly, 2015) can provide guidance on how to understand these concepts in the realm of algorithmic bias in nuanced, context-sensitive ways that are attentive to issues of intersectionality. Works in feminist social epistemology and philosophy of science can further uncover how these concepts are implicated in measures and methods that are routinely used in algorithm design and evaluation (Johnson, in press).

In addition, the harms of algorithmic bias are often treated in a relatively narrow way. Philosophical research on algorithmic bias has centered on concerns about distributive or allocative harms, leaving open foundational questions about the potential harms of biased and stereotypical representations (as characterized

by Zheng 2016) in contexts such as image classification and natural language processing (see also Eubanks, 2018 on algorithm-induced stigmatization and stereotyping). Broadening our scope, the dynamic, long-term impacts of algorithms are complicated by what Hacking (1995) calls 'looping effects' on individuals and societies (see also Noble, 2018). Philosophical research can inform our understanding of these effects; for example, research in aesthetics can illuminate potential risks in the use of algorithms in the recommendation and curation of artworks.

6 | CONCLUSION

Predictive algorithms are quickly pervading our lives and societies. Organizations adopt algorithms to guide the allocation of benefits and burdens in education, healthcare, workforce, criminal justice, finance, and more. Algorithmic recommendations shape what we experience, and they affect our social ties and affiliations from romance to politics. In many cases, these changes to our personal, institutional, and societal dynamics are accepted only because the pervasive impact of adopting predictive algorithms remains implicit (cf. Winner, 1980). Close philosophical engagement is key to characterize, identify, and respond to harmful biases that can result from the simplistic adoption and use of predictive algorithms in our societies.

ACKNOWLEDGMENTS

Sina Fazelpour was partially supported by the Social Sciences and Humanities Research Council of Canada (PDF #756-2019-0289). Thanks to two anonymous reviewers for valuable feedback on an earlier draft of this paper. Significant portions of this paper were written while Sina Fazelpour was at Carnegie Mellon University.

CONFLICTS OF INTEREST

None.

ORCID

David Danks https://orcid.org/0000-0003-4541-5966

ENDNOTES

[1] This distinction helps to clarify the common situation in which an algorithm with problematic predictions is nonetheless defended as "unbiased" (as with the infamous COMPAS algorithm). Almost always, the algorithm proponent focuses on the algorithm being *statistically* unbiased, while the algorithm opponent focuses on *moral* biases. Both descriptions can be correct when training data include morally problematic phenomena.

[2] This problem is a specific instance of sample selection bias, where inclusion in the dataset depends partly on the values of other variables. All responses rely on strong substantive assumptions (see De-Arteaga et al., 2018).

[3] Important value judgments arise in this train-test split (e.g., with respect to distributional properties of the two datasets).

[4] Developers sometimes seem less cognizant about these value dimensions, such as when success criteria involve interpretability. In these cases, these metrics (e.g., simplicity) do not necessarily correspond to the actual aims of decision-makers that drive demands for interpretability (e.g., trust, decision support, etc.). For a thoughtful discussion, see Krishnan (2020).

[5] "When a measure becomes a target, it ceases to be a good measure."

[6] For a thorough critique of gender- and race-blind social policies (beyond algorithm-based ones), see Anderson, 2010.

[7] We omit the particular mathematical translations of each of these concepts, as they are widely available in the fair ML literature.

[8] Many of these insights involve a re-discovery of similar results in literature on test fairness; see the history provided in Hutchinson and Mitchel (2019).

[9] Some U.S. law also supports causal metrics, as judgments of employment discrimination can turn on "whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had remained the same" (Carson v. Bethlehem Steel Corp. 82 F .3d 157 159 (7th Circ. 1996)).

REFERENCES

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, *7*(2), 2053951720949566.

Anderson, E. (2010). *The imperative of integration*. Princeton University Press.

Antony, L. (2016). Bias: Friend or foe? Reflections on Saulish skepticism. *Implicit Bias and Philosophy*, *1*, 157–190.

Austin, N. L., Carter, R. T., & Vaux, A. (1990). The role of racial identity in Black students' attitudes toward counseling and counseling centers. *Journal of College Student Development*.

Barabas, C., Virza, M., Dinakar, K., Ito, J., & Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pp. 62–76. New York City: ACM.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved from http://www.fairmlbook.org

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Beeghly, E. (2015). What is a stereotype? What is stereotyping? *Hypatia*, *30*(4), 675–691.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Wiley.

Berendt, B. (2019). AI for the common good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn. Journal of Behavioral Robotics*, *10*(1), 44–65.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Journal of Machine Learning Research*, *81*, 1–15.

Carton, S., Mahmud, A., Cody, C., Helsby, J., Park, Y., Patterson, E., Ghani, R., Joseph, K., Walsh, J., & Haynes, L. (2016). Identifying police officers at risk of adverse events. In Proceedings of KDD 2016. San Francisco.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. Sydney, NSW: ACM.

Castro, C. (2019). What's wrong with machine bias. *Ergo*, *6* (15).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82–89.

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv preprint arXiv:1808.00023.

Corbett-Davies, S., Goel, S., & González-Bailón, S. (2017). Even imperfect algorithms can improve the criminal justice system. *New York Times*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. Halifax, Canada: ACM.

Courtland, R. (2018). Bias detectives: the researchers striving to make algorithms fair. *Nature*, *558*(7710), 357–360.

Coyle, D., & Weller, A. (2020). "Explaining" machine learning reveals policy challenges. *Science*, *368*(6498), 1433–1434.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, *87*(4), 568–589.

Dai, J., Fazelpour, S., & Lipton, Z. C. (2021). Fair machine learning under partial compliance. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Honolulu: ACM.

Danks, D. (2019). The value of trustworthy AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 521–522. Honolulu: ACM.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 4691–4697. Melbourne, Australia.

De-Arteaga, M., Dubrawski, A., & Chouldechova, A. (2018). *Learning under selective labels in the presence of expert consistency*. arXiv preprint arXiv:1807.00905.

De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. Honolulu: ACM.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc*, *7*(7.4), 1.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114, 126.

Dotan, R. (2020). Theory choice, non-epistemic values, and machine learning. *Synthese*. 1–21. https://doi.org/10.1007/s11229-020-02773-2

Douglas, H. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press.

Dutta, S., Wei, D., Yueksel, H., Chen, P., Liu, S., & Varshney, K. (2020). Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. In *Proceedings of the 37th International Conference on Machine Learning, in Proceedings of Machine Learning Research*, (Vol. 119, pp. 2803–2813). Retrieved from http://proceedings.mlr.press/v119/dutta20a.html

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. Cambridge, MA: ACM.

Dwork, C., Ilvento, C., & Jagadeesan, M. (2020). Individual Fairness in Pipelines. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pp. 160–171. New York City: ACM.

Eubanks, V. (2018). *Automating inequality*. St. Martin's Press.

Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 57–63. New York City: ACM.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. Sydney: ACM.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

Gendler, T. S. (2011). On the epistemic costs of implicit bias. *Philosophical Studies*, *156*, 33–63.

Glymour, B., & Herington, J. (2019). Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 269–278. Atlanta: ACM.

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 90–99. Atlanta: ACM.

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, A. J. Premack (Eds.) *Causal cognition: A multidisciplinary debate*. Clarendon Press.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Ben Hadj Hassen, A., Thomas, L., Enk, A., Uhlmann, L., & Reader study level-I and level-II Groups. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists, *Annals of Oncology*, *29*(8), 1836–1842.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*, 3315–3323.

Hart, H. L. A. (1961). *The concept of law*. Oxford University Press.

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*, 811.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915.

Hu, L., & Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1389–1398. Geneva: ACM.

Hu, L., & Kohler-Hausmann, I. (2020). What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513. Barcelona: ACM.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 49–58). Atlanta: ACM.

Johnson, G. (2020). Algorithmic bias: On the implicit biases of social technology. *Synthese*, 1-21.

Johnson, G. (In press). Are algorithms value-free? Feminist theoretical virtues in machine learning. Journal of Moral Philosophy.

Kallus, N., & Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*, Stockholm.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems 33*(1), 1–33.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, *30*, 656–666.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018a). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*(1), 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018b). Discrimination in the age of algorithms. *Journal of Legal Analysis*, *10*, 113–174.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*, 487–502.

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, *578*(7793), 34–36.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). *Counterfactual fairness*. arXiv preprint arXiv:1703.06856.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284. Halifax, Canada: ACM.

Leben, D. (2020). Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 86–92. New York City: ACM.

Liao, S., & Huebner, B. (In press). Oppressive things. *Philosophy and Phenomenological Research*.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. Stockholm: PMLR.

Loftus, J. R., Russell, C., Kusner, M. J., & Silva, R. (2018). *Causal reasoning for algorithmic fairness*. arXiv preprint arXiv:1805.05859.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21.

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591.

Mayson, S. G. (2018). Bias in, bias out. *Yale Law Journal*, *128*, 2218.

Merry, S. E. (2016). *The seductions of quantification: Measuring human rights, gender violence, and sex trafficking*. University of Chicago Press.

Milli, S., Miller, J., Dragan, A. D., & Hardt, M. (2019). The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239. Atlanta: ACM.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Muldoon, R. (2016). *Social contract theory for a diverse world: Beyond tolerance*. Taylor & Francis.

Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* pp. 39–48. Atlanta: ACM.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, *4*(7), 736–745.

Pildes, R. H., & Anderson, E. S. (1990). Slinging arrows at democracy: Social choice theory, value pluralism, and democratic politics. *Columbia Law Review*, *90*(8), 2121–2214.

Rambachan, A., & Roth, J. (2020). Bias in, bias out? evaluating the folk wisdom. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68).Atlanta: ACM.

Simon, J. (2017). Value-sensitive design and responsible research and innovation. In S. Ove Hansson (Ed.), *The ethics of technology*, 219–236. London, UK: Rowman & Littlefield.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330.

Tuana, N. (2010). Leading with ethics, aiming for policy: New opportunities for philosophy of science. *Synthese*, *177*(3), 471–492.

Véliz, C. (2020). *Privacy is power: Why and how you should take back control of your data*. Random House.
Westen, P. (1982). The empty idea of equality. *Harvard Law Review*, 537–596.
Winner, L. (1980). Do artifacts have politics? *Dædalus*, 121–136.
Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. Canary Islands: PMLR.
Zhang, J., & Bareinboim, E. (2018). Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
Zheng, R. (2016). Why yellow fever isn't flattering: A case against racial fetishes. *Journal of the American Philosophical Association*, *2*(3), 400–419.

AUTHOR BIOGRAPHIES

Sina Fazelpour is Assistant Professor of Philosophy and Computer Science at Northeastern University. He was previously a Social Sciences and Humanities Research Council (SSHRC) Postdoctoral Fellow at the Department of Philosophy at Carnegie Mellon University. He holds a PhD in Philosophy from the University of British Columbia, a M.Sc in medical biophysics from the University of Toronto, and a B.Eng in electrical and biomedical engineering from McMaster University. His primary research focus is on issues of reliability and accountability in predictive and decision-making algorithms. He also works on understanding the impact of diversity on group dynamics and performance. He has published in the philosophy of science, cognitive science and ethics of artificial intelligence, and his research has been supported by Joseph-Armand Bombardier Canada Graduate Scholarship, the Block Center for Technology and Society, the Templeton Foundation, and Natural Sciences and Engineering Research Council of Canada.

David Danks is L.L. Thurstone Professor of Philosophy & Psychology, and Head of the Department of Philosophy, at Carnegie Mellon University. He is also the Chief Ethicist of CMU's Block Center for Technology & Society; and co-director of CMU's Center for Informed Democracy and Social Cybersecurity (IDeaS). His research interests are at the intersection of philosophy, cognitive science, and machine learning, including the ethical, psychological, and policy issues around AI and robotics in transportation, healthcare, privacy, and security. He is the author of *Unifying the Mind: Cognitive Representations as Graphical Models* (2014, The MIT Press). Danks is the recipient of a James S. McDonnell Foundation Scholar Award, as well as an Andrew Carnegie Fellowship. He received an A.B. in Philosophy from Princeton University, and a Ph.D. in Philosophy from University of California, San Diego.

How to cite this article: Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760