Informational Richness and its Impact on Algorithmic Fairness

(forthcoming in *Philosophical Studies*)

Marcello Di Bello[1] and Ruobin Gong[2†]

[1]School of Historical, Philosophical and Religious Studies, Arizona State University, 975 S. Myrtle Ave P.O. Box 874302, Tempe, 85287, AZ, United States.

[2]Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Hill Center 404, Piscataway, 08854, New Jersey, Unite States.

Contributing authors: mdibello@asu.edu; ruobin.gong@rutgers.edu;
[†]These authors contributed equally to this work.

Abstract

The literature on algorithmic fairness has examined exogenous sources of biases such as shortcomings in the data and structural injustices in society. It has also examined internal sources of bias as evidenced by a number of impossibility theorems showing that no algorithm can concurrently satisfy multiple criteria of fairness. This paper contributes to the literature stemming from the impossibility theorems by examining how informational richness affects the accuracy and fairness of predictive algorithms. With the aid of a computer simulation, we show that informational richness is the engine that drives improvements in the performance of a predictive algorithm, in terms of both accuracy and fairness. The centrality of informational richness suggests that classification parity, a popular criterion of algorithmic fairness, should be given relatively little weight. But we caution that the centrality of informational richness should be taken with a grain of salt in light of practical limitations, in particular, the so-called bias-variance trade off.

Keywords: Algorithmic Fairness, Classification Parity, Predictive parity, Impossibility theorems, Computer simulation, Informational richness, Conscientiousness, Bias-variance trade off

1 Introduction

Many decisions matter for people's lives: whether an applicant is granted a loan; whether a patient is given medical care; whether a defendant is placed in preventative detention. Human beings often make these decisions: bankers, doctors, judges. But their judgment can be mistaken. An applicant in good financial standing could be denied a loan; a sick patient could be denied treatment; someone who is not going to commit a crime could be put in jail. Historically, these mistakes have adversely impacted select social groups the most, racial and gender minorities, and the economically worse off.[1] If human judgment is aided by machine judgment, the accuracy and fairness of our decisions would improve—or so some argue.[2] Data about every aspect of our lives are now easily available. Individuals can be classified based on the predictive attributes they possess and assigned a risk score expressing the probability that the outcome of interest will occur. Risk scores are not infallible, of course. But—so the argument goes—since they are based on data, predictive algorithms should perform better than human judgment in both accuracy and fairness.

Despite this optimism, many are alarmed. Three reasons for concern exist in the literature. First, many worry about the ripple effects of the historical data on which predictive algorithms are trained. Defects in the data can have far reaching, harmful consequences.[3] Call this the *distorted data argument.* Second, even if the training data were not distorted and portrayed an accurate picture, the worry about predictive algorithms would not necessarily subside. Trends in the data reflect trends in society. The society we are in is replete with group disparities in wealth, crime, health.[4] If—as many have argued—the status quo is shaped by structural injustice against historically disadvantaged groups, an accurate prediction would reinforce an unjust reality.[5] This is especially the case when algorithmic predictions prompt a punitive or coercive decision such as loan rejection or preventative detention. Call this the *historical injustice argument.*[6]

[1] There is strong evidence of an association between race and differential treatment by health care providers (McKinlay, 1996; Schulman et al., 1999; Chen et al., 2001; Petersen et al., 2002). Whether or not these differences are explained by implicit biases is unclear (Dehon et al., 2017). On lending practises, there is a growing body of literature documenting the impact of redlining on economic inequalities today (Aaronson et al., 2021; Ladd, 1998). The justice system is filled with racial disparities at different stages (Rehavi and Starr, 2014; Gross et al., 2022).

[2] For example, the American Civil Liberty Union of New Jersey argued that the deployment of predictive algorithms in criminal justice can end the unfair system of bail that most disproportionately harms the poor; see *https://www.aclu-nj.org/theissues/criminaljustice/pretrial-justice-reform.* For a more detailed defense of this claim, see Slobogin (2021). The consulting firm McKinsey estimated that predictive algorithms can save $300 billion every year in U.S. healthcare costs (Manyika et al., 2011). More generally, for the positive impact of big data in health care, see (Raghupathi and Raghupathi, 2014).

[3] Data can be defective because of their reliance on *proxies*, for example, when arrest data are used as proxies for actual criminal offending (Barabas et al., 2019) or when healthcare costs are used as proxies for actual medical needs (Obermeyer et al., 2019). Beside the proxy problem (also known as measurement problem), biases can arise during data collection, for example, when certain groups are under-sampled. For an overview of sources of bias in the data, see, among others, Suresh and Guttag (2021). For an analysis of the implications of biased data from the standpoint of US constitutional law, see Barocas and Selbst (2016).

[4] Along similar lines, Mitchell et al. (2021) draw a distinction between statistical bias (a mismatch between the world and the sample used to train the model) and societal bias (a mismatch between the world as it is and the world as it should be).

[5] Define "structural injustice" as any historically entrenched distribution of goods, benefits, powers and advantages (or their negative correlates) among social groups, where such distribution negatively impact the well-being of specific social groups and not others. See Powers and Faden (2019) and Young (2003).

[6] Deborah Hellman (2021) calls this phenomenon *compounding injustice.* Facts grounded in past injustices are used as the basis for making punitive decisions in the present, thereby compounding the past injustice.

These two arguments point to exogenous problems that lie on the input side of things: the data used to train predictive algorithms and the unjust society from which data originate. One might conjecture that, if distortions in the data and injustices in society were eliminated, predictive algorithms should no longer be cause for concern. But this conjecture would be premature. Predictive algorithms can still be the target of what we might call an *inner critique*. This inner critique stems from a number of theorems in the computer science literature about the impossibility of algorithmic fairness. It is this inner critique that we focus on in this paper.

To state the impossibility theorems, we should begin with laying out a number of fairness criteria of algorithmic performance. These criteria are an attempt to capture in formal, mathematical language the requirement that a predictive algorithm should treat people fairly. Predictive parity and classification parity are two of the most common criteria in the computer science literature. Predictive parity requires that the rate at which an algorithm's predictions are correct—the fraction of predictions that are correct—be the same across groups. For example, predictive parity would be violated whenever people that the algorithm predicted would default on their loan ended up actually defaulting, say, in 90% of the cases if they were white, but only in 60% of the cases if they were black. Another popular criterion, classification parity, requires that the rates at which individuals are the recipients of correct predictions be the same across groups. If the prediction and the outcome are both binaries, this criterion requires that the true positive and true negative rates—the fraction of truly positive people that are predicted to be positive, and the fraction of truly negative people that are predicted to be negative—be the same across social groups. It would be a violation of classification parity if people who were, in an objective sense, not going to default on their loan were erroneously predicted to default in 10% of the cases if they were white and 30% if they were black. Stated more formally, the two criteria require the equalization—across two distinct groups of interest—of two different conditional distributions. Predictive parity requires the equalization across groups of the conditional distributions of the outcome of interest given the prediction.[7] Instead, classification parity requires the equalization across groups of the conditional distributions of the prediction given the outcome of interest.[8]

[7] Formally, the algorithm's prediction should satisfy the following equality between conditional probability statements:

$$P\left(Y=1 \mid S \geq a, G=g\right)=P\left(Y=1 \mid S \geq a, G=g^{\prime}\right) \quad \forall g \neq g^{\prime},$$

where Y is the binary outcome to be predicted (which can take values 1 or 0) and G is the group membership based on a protected classification. The expression $S \geq a$ is the algorithm's binary prediction of the positive outcome $Y = 1$. Predictive algorithms usually make a fine-grained prediction in terms of a risk score S. The greater the score, the greater the probability of the outcome. The algorithm's binary prediction results by thresholding the risk score at some value a that is considered sufficiently high. Another common criterion of predictive parity is calibration, a more fine-grained version of equal positive predictive value. This measure of fairness is not dependent on a decision threshold. It compares the predictive accuracy of the algorithm across groups for each risk score, not just risk scores above the threshold. A predictive algorithm is *relatively calibrated* (Chouldechova, 2017; Corbett-Davies and Goel, 2018) if

$$P\left(Y=1 \mid S, G=g\right)=P\left(Y=1 \mid S, G=g^{\prime}\right) \qquad \forall g \neq g^{\prime}.$$

If the risk score further satisfies $P\left(Y=1 \mid S, G\right)=S$, we say that it is *absolutely calibrated* (Kleinberg et al., 2017).

[8] Formally, a predictive algorithm satisfies equal classification accuracy if it has the same false positive rates across groups:

$$P\left(S \geq a \mid Y=0, G=g\right)=P\left(S \geq a \mid Y=0, G=g^{\prime}\right) \quad \forall g \neq g^{\prime},$$

Predictive and classification parity are group measures of fairness. They require that differences in algorithmic performance across groups be eliminated.[9] They are plausible measures of algorithmic fairness on the assumption that differences in algorithmic performance will eventuate in differences in the allocation of benefits and burdens across groups since algorithmic predictions guide these allocations. But the major obstacle toward satisfying these criteria is constituted by a number of impossibility theorems, now well-known in the computer science literature. These theorems show that no algorithm can satisfy all candidate criteria of algorithmic group fairness, in particular, no algorithm can satisfy both predictive and classification parity.[10] These theorems only require minimal assumptions: first, the algorithm can make mistakes; second, the two groups being compared have different base rates of the outcome of interest, say different rates of loan defaulting. Exogenous factors, such as distorted data or historical injustice, cannot be blamed since the impossibility theorems are a mere mathematical consequence of the fact that the conditional distributions of two uncertain quantities—the probability of the outcome given the prediction or the probability of the prediction given the outcome—are generally untethered. In this sense, the impossibility of concurrently satisfying different fairness criteria can be viewed as an inner critique of predictive algorithms.[11]

Reactions to the impossibility theorems have been threefold. One line of argument emphasizes pragmatic considerations. Many in the computer science literature have pointed out that algorithmic decisions must confront trade-offs, first between accuracy and fairness but also between the different criteria of fairness themselves.[12] Whether one criterion of fairness takes precedence over another may depend on matters of

as well as the same false negative rates across groups:

$$P\left(S \leq a \mid Y=1, G=g\right)=P\left(S \leq a \mid Y=1, G=g^{\prime}\right) \quad \forall g \neq g^{\prime}.$$

The satisfaction of these conditions depends on a specific risk threshold that is considered high enough to make a positive classification. Balance is another measure of classification parity which, however, does not depend on selecting a specific risk threshold. A predictive algorithm is said to be balanced if it assigns on average the same risk scores for people with the same positive outcome ($Y=1$) or negative outcome ($Y=0$) in each group membership. In terms of expectation, balance can be defined as follows:

$$E\left(S \mid Y=y, G=g\right)=E\left(S \mid Y=y\right)$$

for any group g and outcome $y=0$ or 1.

[9] In contrast, individual fairness is often understood as equal treatment of similarly situated individuals (Dwork et al., 2012; Sharifi-Malvajerdi et al., 2019). This conception of algorithmic fairness tracks how an individual is treated relative to others by constructing a counterfactual (Kusner et al., 2018). On the apparent conflict between individual and group fairness, see Binns (2020).

[10] The two most well-known impossibility results are due to Chouldechova (2017) and Kleinberg et al. (2017). An earlier result was proven by Borsboom et al. (2008). There is also a possibility result due to Reich and Vijaykumar (2021) who show that classification parity (specifically, equal false positive and false negative rates across groups) and predictive parity (specifically, calibration) can be concurrently satisfied.

[11] Some claim that different performance criteria of algorithmic fairness embody different moral commitments about what fairness requires. In this sense, the impossibility theorems underscore a conflict between different moral commitments about algorithmic fairness (Heidari et al., 2019). This interpretation is compatible with our own. Our claim that the impossibility theorems constitute an inner critique underscores the fact that violations of fairness criteria can occur absent *exogenous* sources of bias in the data or in society. On a more technical level, a popular explanation for why these violations of fairness criteria occur even without exogenous biases appeals to the so-called problem of *infra-marginality.* As soon as two groups have differences in prevalence—say, differences in criminality, financial stability or health—the shape of the risk distributions of the two groups, as viewed by the predictive algorithm, will be different. This implies, inevitably, that the rate of correct predictions will differ across groups, thus giving raise to violations of one criterion of fairness or another (Corbett-Davies and Goel, 2018).

[12] On trade-offs between different fairness criteria, see Berk et al. (2021) and Lee et al. (2021).

context.[13] A second line of argument rejects altogether the dilemma raised by the impossibility theorems and emphasizes the goal of realizing substantive fairness, as well as ending oppression and historical injustice.[14] Finally, a third line of argument is conceptual and is prevalent in the philosophical literature. This approach resists predictive or classification parity as adequate criteria of algorithmic fairness because they do not genuinely capture requirements of fairness. This resistance is justified by constructing hypothetical scenarios in which our intuitions about algorithmic fairness (or unfairness) diverge from the satisfaction (or violation) of a fairness criterion of interest.[15]

Our contribution adds to each of these lines of argument, although our focus is on the conceptual point. We agree that common criteria of algorithmic fairness do not fully capture what it means for predictive algorithms to be fair. But, besides offering hypothetical scenarios as counterexamples, the current literature does not explain, in a principled manner, why these fairness criteria fall short. Methodologically, reliance on hypothetical scenarios can also be questioned insofar as these scenarios are not representative of how algorithmic predictions are made. To remedy this, we construct a more realistic probabilistic model designed to mimic how predictive algorithms are trained on data in which group membership is causally implicated. We then examine the model via a simulation study.

Another limitation of the existing literature on the impossibility theorems is its primary focus on what we call criteria of *performance*. These criteria track algorithmic performance in the long run: they track how often an algorithm makes mistakes (accuracy) and how these mistakes are distributed across groups (fairness). But besides performance, another dimension deserves attention, what we call *conscientiousness*. Compare a doctor who makes diagnoses on just few sparse symptoms and a doctor who carefully assesses all the relevant symptoms that a patient exhibits. By taking into account more information, the second doctor is more conscientious than the first. Similarly, an algorithm can base its predictions on a richer or poorer set of predictive features. The richer the information, the more conscientious the predictions.[16]

[13] On the contextuality of criteria of algorithmic fairness within a theory of justice that applies to predictions, as opposed to decisions, see Lazar and Stone (ms).

[14] On this more radical approach, see Green (2022).

[15] In philosophy, Brian Hedden (2021) and Robert Long (2021) have provided the most discussed examples.

[16] The idea of conscientiousness has been discussed—under different names—in both the philosophical and computer science literature in different ways. In the philosophical literature, the idea of conscientiousness is closely related to what some call "the right to be treated as an individual." This right can be understood in an informational sense, roughly as the right to be judged on as much relevant information as what is reasonably available (Lippert-Rasmussen, 2011). Others have emphasized the imperative of avoiding doxastic negligence and collecting more information if appropriate (Zimmermann and Lee-Stronach, 2022). Another, non-informational conceptions of the right to be treated as an individual focuses on the fair allocation of risks and burdens (Castro, 2019; Jorgensen, 2022). In the computer science literature, some have suggested that further screening or collecting more data about select groups can improve the fairness performance of predictive algorithms (Chen et al., 2018; Cai et al., 2020). We are sympathetic with these approaches, but our analysis differs in two ways. First, we are not advocating that only select groups be subject to further screening or data gathering as this may increase surveillance of already marginalized communities. Second, we are interested in examining how conscientiousness impact the different performance criteria of algorithmic fairness. As we will see, improvements in conscientiousness do not impact all performance criteria of fairness equally (Section 3). This observation will then be the basis for an argument against classification parity (Section 5).

Our focus on conscientiousness paired with the simulation study will help to see that not all performance measures are born equal. Under normal circumstances, accuracy and conscientiousness go hand in hand, and fairness—understood as predictive parity—does too. Classification parity is the outlier, and this makes it a particularly objectionable criterion of algorithmic fairness. On the other hand, all performance measures are prone to manipulation: they can be violated or satisfied by means of *ad hoc* manipulations of the characteristics of the groups being compared. In such cases, performance measures fail to align with the intuitive requirements of algorithmic fairness. We will argue that this failure is explained by the extent to which performance measures, such as classification or predictive parity, deviate from conscientiousness.

Our contribution also helps to clarify the pragmatic point about trade-offs. It is sometimes asserted that there is a tension between accuracy and fairness: an improvement in accuracy can be detrimental for fairness.[17] In addition, the impossibility theorems mentioned earlier demonstrate that a tension exists within fairness itself, among different criteria of fairness. If one wanted to satisfy all performance measures of algorithmic fairness, trade-offs will be inevitable. But the extent of this inevitability must not be exaggerated. If—as we will demonstrate—accuracy, conscientiousness and predictive parity go together, and classification parity is the outlier, the trade-offs between different measures of fairness as well as between fairness and accuracy become less pressing.

Our paper does not directly address questions of historical injustice and how the latter should inform our theorizing about algorithmic fairness. It is an under-explored topic in the literature to what extent performance criteria such as classification and predictive parity reflect inequalities in society.[18] This relationship is unlikely to be straightforward, however.[19] More work certainly needs to be done, but our simulation study shows that the group for which the violation of a performance criterion is most detrimental is not fixed in advance. For example, a higher rate of false loan rejections may affect the group with higher prevalence of loan default or the group with a lower prevalence, where one or the other may be the disadvantaged group. So, given this variability, violations of algorithmic fairness criteria need not reflect in a systematic way patterns of structural inequalities across groups in society.

The plan is as follows. Section 2 provides the technical and conceptual backdrop of our investigation, specifically, the contrast between idealized and empirical risk. Section 3 describes the probabilistic model and the computer simulation. Section 4 argues that performance criteria of fairness cannot be divorced from questions of

[17] On the trade-off between accuracy and fairness, see Menon and Williamson (2018). Kearns and Roth (2019) discuss the concept of a Pareto frontier between accuracy and fairness.

[18] The literature on causal criterion of algorithmic fairness has begun to address these questions, see e.g. Chiappa and Gillam (2018).

[19] For one thing, group differences in prevalence—which drive in part violations of predictive and classification parity—are not necessarily due to structural injustice. There exist several layers of inequality that may exist in society. Some inequalities are certainly due to structural, historical injustices and discrimination, but others may be less pernicious and due to differences in preferences or priorities among groups (Lee et al., 2021). At the same time, violations of fairness performance criteria could still cause harm even without historical conditions of structural injustice. Consider two communities whose wealth happens to be different, but not for reason of structural injustice. If a community experiences, say, a higher rate of false loan rejections, this difference in the long run may entrench their economic disadvantage. Or suppose the algorithm's predictive accuracy is worse for one community compared to another. This will have negative reputational costs, for example, if one community is viewed as less capable of repaying loans.

conscientiousness. Section 5 argues that classification parity, a popular performance measure of algorithmic fairness, has limited significance. Section 6 discusses some complications, both conceptually and practically. Section 7 situates the notion of conscientiousness within the broader distinction between performance criteria and attitudinal criteria of algorithmic fairness.

2 Individual risk and its estimation

To predict an unknown fact about an individual, the decision maker who does not rely on their own hunches and intuitions can take advantage of a predictive algorithm, also called—perhaps more appropriately—*risk model.* A risk model or predictive algorithm is an abstract, evidence-based representation of the correlations between certain features (attributes, traits, predictors) an individual may possess and an outcome of interest. To understand how predictive algorithms can make mistakes and why their performance can differ across groups, we pry open this conceptual construction to examine its inner workings. We focus in particular on the degree of informational richness that is the basis of the algorithmic estimation of the risk ascribed to individuals.

We think of each individual as characterized by an infinite collection of measurable attributes, features or traits, denoted as $\vec{X}_\infty = \{X_1, X_2, \ldots\}$. This infinite collection encompasses all information that ever exists about this person, including demographic, genetic, behavioral and psychological data at any given time point. This information can be so detailed to uniquely characterize an individual. That is, knowing $\vec{X}_\infty$ is equivalent of knowing the individual, and indeed we will denote the individual as $\vec{X}_\infty$. The unknown binary outcome we wish to predict about the individual is denoted by Y. For example, Y may denote whether an applicant will default on their loan ($Y = 1$) or not ($Y = 0$).

Presumably, there exists an objective relationship between an individual's attributes $\vec{X}_\infty$ and the outcome Y. This relationship could in principle be captured by S_∞, the *idealized risk score* of the individual. More precisely, the idealized risk score is denoted by

$$S_\infty \left(\vec{X}_\infty; \theta^*_\infty \right),$$

where $\theta^*_\infty = (\theta^*_1, \theta^*_2, \ldots)$ is the ideal value of the (possibly infinite-dimensional) parameter, which governs in the finest detail the relationship between the idealized risk and the infinite set of attributes $\vec{X}_\infty$. The notation emphasizes the functional dependence of the idealized risk score on both $\vec{X}_\infty$ and θ^*_∞. In contexts where this dependence is not important, we simply write S_∞ for short.

The idealized risk score (idealized risk, for short) is the *best* probabilistic description of the individual's outcome Y. Once the value of the individual's idealized risk has been learned, no additional information can be more indicative about the unknown value of Y. There is a distinction between two ways through which the modeler may conceptualize the meaning of the best probabilistic description, one *deterministic* and one *stochastic.* On the first conceptualization, the idealized risk S_∞ is thought to track the outcome Y perfectly. That is, S_∞ is either zero or one, and $Y = S_\infty$. Had it been possible for the modeler to access the infinitely rich information about the individual, the modeler would know the outcome for sure. By contrast, the stochastic conceptualization stipulates that while there is nothing more that can be learned about Y beyond S_∞, knowing its value still does not allow us to pin down the outcome with certainty. In other words, there is some randomness in the individual's outcome that just cannot be fully tamed. Under this conceptualization, S_∞ is a probability whose value is anywhere between zero and one.[20] For most of the paper, we will assume the deterministic conceptualization, but discuss the implications of the stochastic one toward the end.

Regardless of the conceptualization of the idealized risk, the infinitely detailed attribute collection $\vec{X}_\infty$ remains a hypothetical construction. In reality, only a finite subset of the content of $\vec{X}_\infty$ can be accessed, say p dimensions of it. Denote the finite accessible information about an individual by $\vec{X}_p$. The dimensionality p may reflect the practical limitation of how much information can be collected about an individual or the modeler's intention to include only certain attributes that are deemed admissible. For all individuals who share the same accessible information $\vec{X}_p$, the modelers supplies an estimated *empirical risk*, denoted by $\hat{S}_p$. More precisely,

$$\hat{S}_p\left(\vec{X}_p; \hat{\theta}_p\right),$$

or $\hat{S}_p$ for short, whenever the dependence on its arguments $\vec{X}_p$ and $\hat{\theta}_p$ may be suppressed. For a binary outcome Y, the empirical risk $\hat{S}_p$ is a fractional number between zero and one, with a larger value suggesting $Y = 1$ as more likely.

To carry out the estimation of individual risk in practice, the modeler must engage in the postulation, fitting, and selection among a collection of candidate risk models. To do so, they must operate within realistic bounds of their domain knowledge, available information, and computation capacities. So, throughout this process, they have several practical choices to make. They must determine the appropriate dimension p, what we will call the *informational richness* of $\vec{X}_p$. They must also determine the functional form of the risk model to be used alongside the input $\vec{X}_p$. Typically, the risk model is assumed to belong to a family of functions, to allow for a good approximation to the idealized individual risk. The family from which the risk model is chosen is capable of capturing increased complexity as the richness of the attribute set $\vec{X}_p$ increases, while commanding a larger parameter space as well. Lastly, the value of the parameter that governs the function must also be determined. The hat notation in $\hat{S}_p$ signifies we are dealing with estimates from the observed data that bear variability due to the data collection process.

We take this conceptual framework to be relatively uncontroversial. The upshot here is that the objective of algorithmic prediction is to approximate, in the best way possible, the idealized individual risk (the objective risk that each individual will do this or that) by means of the empirical individual risk (the risk that each individual will do this or that based on the information and modeling assumptions available).

In light of this conceptual framework, it is instructive to revisit an example by Brian Hedden (2021). This example is part of an argument against nearly all performance

[20] What's implied of the relationship between Y and S_∞ is also weaker than the equality relationship, but one reasonable requirement is that the idealized risk satisfies absolute calibration: $P(Y = 1 \mid S_\infty) = S_\infty$, where the probability P reflects the untamed randomness inherent in the outcome Y.

criteria of algorithmic fairness. Hedden describes a scenario in which each person is given a biased coin, reflecting their objective risk (say, for concreteness, the person's objective risk of committing a crime). Suppose a predictive algorithm can faithfully track the objective risk of each person and in this way can assign an equivalent risk score. There is no mismatch between objective risk and the algorithm's risk score. By thresholding the risk score at some value, the algorithm can make binary predictions about the outcome. But now suppose people are sorted into two rooms, and it just so happens that the distributions of the objectives risks (and thus of the algorithm's risk scores) across the two rooms are different. As a consequence, the algorithm's rates of false positives and false negatives across the two rooms—so long as the same threshold is used—will differ. Classification parity is thus violated. Since the objective risk distributions differ across the two rooms, predictive parity will also be violated for analogous reasons.[21] Hedden points out that such violations of classification and predictive parity do not make the algorithm unfair. Hence, these criteria cannot be criteria of fairness.

Some have objected to this argument because it is artificial and removed from the practice of algorithmic predictions.[22] We think this criticism is well-founded, but we also agree with Hedden that the algorithm in the scenario is intuitively fair. Why, exactly, should we think so? Hedden appeals to our intuitions. But there is a more principled answer: the predictive algorithm in the example does what it is supposed to do in the best way possible and does that equally well across every single individual. The algorithm has all the information it can possibly have about each individual, and that information is contained in a perfect approximation of the objective bias of the coin, what we called the idealized individual risk. So the empirical risk is the same as the idealized risk, for each individual. The predictive algorithm can do no better. That is why we judge it to be fair.

Still, there is no denying that Hedden's scenario has limited significance because it is artificial. There are at least two reasons for that. First, the distribution of the coin biases just so happens to be different across groups and is assumed to be causally irrelevant. That the distribution of the biases is different across the two groups—different rooms in Hedden's story—is key to bring about violations of most fairness criteria, but is also irrelevant for everything else.[23] In reality, the distributions of predictive features will differ across groups, and group membership is often causally implicated

[21] Hedden's argument does not assume that the people in the two rooms have different base rates. The distribution of their risks is assumed to be different, however. This fact then triggers a violation of the performance criteria of fairness. This is a consequence of the problem of inframarginality; see footnote 11.

[22] For a more extensive critique, see Vigano' et al. (2022).

[23] Another scenario in the philosophical literature, due to Robert Long (2021), makes a similar assumption. Suppose you are an undergraduate student in a large course. For the purpose of grading your homework, you could be assigned to section 1 or section 2. Homework is graded exactly in the same way in the two sections, but it just so happens that the base rate of true A papers is higher in section 1 than in section 2. If the predictive accuracy of the grades is the same across sections, the rate at which true A papers are correctly graded will differ across the two sections. So there will be a disparity in classification errors across the two sections. But, Long argues, this disparity should not raise fairness concerns. Suppose, for concreteness, that true A papers in section 2 are incorrectly graded more often than in section 1. It would be odd for a student in section 2 to complain they were unfairly treated because true A papers were incorrectly graded in section 2 more often. Had the student been in section 1, they would have been graded in the same way. They would have gotten the same grade since being in one section or another is irrelevant for how students are graded. The counterfactual hold simply because group membership is causally irrelevant.

in this difference. Second, no algorithm can perfectly approximate the idealized individual risk and thus algorithms will usually rely on the empirical individual risk. Our discussion in what follows will remove the sources of artificiality just identified. We will ask the following question: how do fairness criteria such as classification and predictive parity perform in more realistic contexts—that is, assuming that (1) groups membership is causally relevant and (2) predictive algorithms rely on the empirical individual risk? The computer simulation in the next section will allow us to address this question.

Before moving on, a clarification about the need of relaxing the idealizations in Hedden's example is in order. One might argue that if a certain criterion of fairness is shown to be inapplicable under idealized conditions, then *a fortiori* the same criterion would be inapplicable under more realistic conditions. But this argument is too quick. Even if—as Hedden has shown—predictive algorithms necessarily violate several performance criteria under idealized conditions and this violation is not intuitively unfair, the same violation under more realistic conditions may still count as unfair. For example, a predictive algorithm whose risk scores perfectly track the objective risks may count as intuitively fair even if it violates predictive parity. And yet, when the algorithm's risk score no longer track the objective risks, the algorithm need not be regarded as intuitively fair. Under more realistic conditions, the violation of classification parity may become morally problematic.[24]

3 The simulation

We now turn to the computer simulation setup to mimic the mathematical setup introduced in the previous section.[25] As already noted, an individual in principle possesses an infinite number of attributes that make up the specific individual they are. For the purpose of illustration, the simulated dataset represents each individual as possessing a finite number of observable and measurable attributes, $(X_1, \ldots, X_{20})$, each taking different numerical values. While each individual is uniquely different in theory, under the assumed model for data generation, two individuals could well possess the same observed attributes due to their finite dimension.

Data generation

In the simulated population of individuals carrying different attributes, some will bring about the action or outcome we are interested in predicting—defaulting on a loan, committing a crime or developing a medical condition—whereas others will not. As a matter of fact, certain combinations of attributes give rise to the outcome of interest, while other combinations do not. To represent this, we presume a generative model for the idealized risk, which associates all the individual's attributes to their outcome. This will be the *oracle* risk model for the simulation. The input to the generative model is a combination of values of the attributes, such as a certain level of income, a certain age, etc. The idealized risk S_∞ is a deterministic function of the finite-dimensional X's for fixed values of the parameter. The function encodes stronger or weaker contributing

[24] For a similar point, see Lazar and Stone (ms).

[25] The R code of the simulation is available with the authors.

relationships between the attributes and the outcome. The generative model is chosen to be

$$S_\infty = \mathbf{1}\left(\text{Probit}^{-1}\left(\beta_0^* + \beta_1^* X_1 + \cdots + \beta_{20}^* X_{20}\right) \geq 0.5\right), \quad Y = S_\infty, \tag{1}$$

where the X's are the input attributes, Y the outcome, and β^*'s the coefficients governing the relative contribution of the attributes towards the uncertain outcome. That is, given a certain combination of values of the attributes, the function outputs 1 or 0, which in turn determines the outcome. Here, the function between the idealized risk and the outcome is assumed to be deterministic: any two individuals possessing the same combination of attributes will either both bring about the outcome (value 1) or not (value 0).[26]

Group disparities

For simplicity, we are assuming there are only two groups we are interested in, labelled generically group 0 and group 1. In the simulation, group membership is not one of the attributes (independent variables) used to generate the outcome. The correlation coefficients associated with the attributes are the same for individuals in both groups. The model generating the outcome is, in this sense, group-blind. Despite that, the prevalence of the outcome of interest (say criminal activity, loan defaulting or medical condition) still differs across groups in the simulation. Even if the process generating the outcome is group-blind, the simulated data show group disparities in the distribution of certain attributes and consequently in the distribution of the outcome of interest. This should not be surprising. In fact, it reflects a familiar pattern. Attributes such as income, education, age may contribute to bringing about a certain outcome. These attributes will also be correlated with protected attributes such as race or gender, even though race or gender need not be directly causally implicated in bringing about the outcome.

To model this setting, some attributes in the simulation depend on group membership, while others do not (Figure 1). Thus, the shape of the distribution of the values of the group-dependent attributes differs by group, while it is the same for the group-independent attributes (Figure 2). And since the distribution of certain attributes is different across groups and the attributes influence the occurrence of the outcome, the two groups have different prevalence rates. So, in the simulation—and unlike Hedden's hypothetical scenario—the difference in prevalence rates, is not a fortuitous fact. It is explained by differences in group-dependent attributes.

Fitting the risk model(s)

Once the simulated data are given, we can make inferences from these data as is usually done in statistics and machine learning. In making inferences, we are attempting to recover to the extent possible the true generative model which, from the input

[26] Hedden in the coin example assumed that the relationship between idealized individual risk (the objective chance or bias of the coin) and outcome was stochastic. Each person was associated with a biased coin and the probability of the outcome was determined by the bias. But predictive algorithms need not be thought as working that way. The relationship between idealized risk and outcome can also be deterministic. Here we assume that the relationship is deterministic but relax this assumption later in the paper.

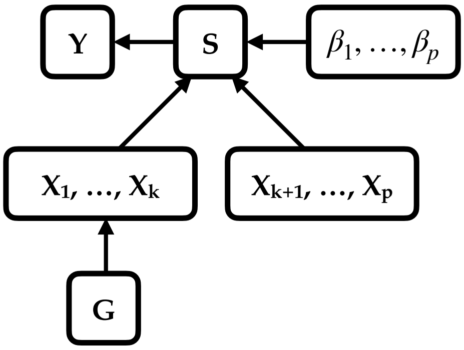

Fig. 1 Graphical representation of the data generating mechanism of the simulation study. The arrows indicate the order of simulation. S stands for the idealized risk score based on all the attributes that make up an individual. The same graph can be used to guide the appropriate specification of the risk model class, in which case S may also stand for the empirical risk score.

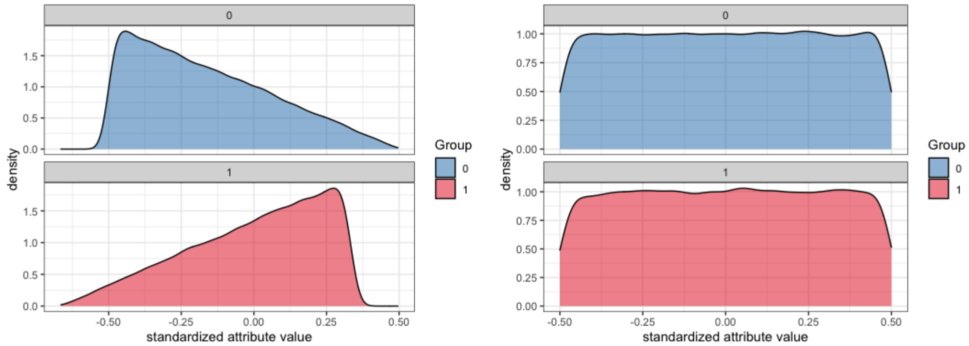

Fig. 2 The empirical distributions of the standardized values of a group-dependent feature (left) and a group-independent feature (right), plotted by group membership.

attributes, returns the outcome via the correlation coefficients. To this end, we use probit regression to create our representation—the *empirical model*—of the true generative model. A common machine learning algorithm, probit regression is naturally suited to our task of predicting a binary outcome, such as defaulting on a loan or committing a crime. We train the probit model on a subset of simulated data, call it the training set. The fitting process finds the optimal coefficients for the model, in the sense that it chooses the parameter value that maximizes a pre-determined objective function.[27] The remaining part of our simulated data will be used to test our model, call it the test or validation set.[28]

[27] For the probit regression model that we examine in this paper, the objective function is simply defined as the data likelihood, rendering maximum likelihood estimation that is guaranteed to consistently and asymptotically efficiently recover the true parameter values in our setting. Other definitions of the objective function, such as those incorporating regularization, may be employed in practice.

[28] To make the simulation more realistic, we vary the composition of group 0 vs group 1 records in the training and the test datasets. In the training set, 60% of the records are from group 0, whereas in the test set, 40% of the records are from group 0. This mimics the possibility that the training and the test sets may over-sample or under-sample some of the groups, so that their sample composition departs from that of the population. Since this variation merely perturbs the group proportion and maintains the ratios between the

In constructing the empirical risk model via probit regression, we considered a number of variations: number of attributes (or predictors); group-dependent vs -independent attributes; size of the training data; and possible mispecifications of the model. We fit a sequence of possible models, some consisting of none or just one true attribute as predictor, all the way to a model consisting of all true attributes. True attributes are those that, as a matter of fact, bring about the outcome of interest in the generative (oracle) model. We also fit a a collection of models, each consisting of a varying number of predictors that are or are not correlated with group, arranged in a different order. We trained our models on training sets of difference sizes. We did not assume that the smaller data set is biased or distorted, only that it has fewer observations. Finally, we fit a collection of models that utilize partially mis-specified predictors.

4 Performance Criteria and Conscientiousness

What do we learn from the simulation? A number of trends emerge. First, informational richness goes hand in hand with predictive accuracy. The more true predictors are used by the risk model, the more accurate the risk model (Figure 3). Accuracy here refers to the "closeness" between the actual value of the individual risk that is assumed to exist, and the algorithm's best judgment of its value.[29] We take informational richness to be an indicator of how many true predictors were used in the risk model.[30] This is a simplification, but is the most salient way to track conscientiousness under the chosen simulation setting. We should note that the seemingly obvious observation that informational richness correlate positively with accuracy shall not be taken for granted. Indeed, an increase in accuracy with increasing data is not automatically obtainable for every risk model, but only for those models that are well-designed and thoughtfully estimated (more on this in Section 6).

The second trend is that, the more accurate the risk model, the better its performance in terms of predictive parity, one of the key formal criteria of algorithmic fairness (Figure 4). Recall that predictive parity requires that the fraction of correct algorithmic predictions be the same across the relevant groups of interest. Thus, we see a strong monotonic trajectory that is common to three different indicators: greater

positive versus negative outcomes within each group identical across the training and the test sets, it does not reflect an outcome-biased sampling scheme and does not constitute an instance of distorted data.

[29] A specification of a loss function is the standard procedure to measure accuracy. The loss function embodies the assessment of closeness between the risk model S and the true outcome Y it is intended to predict. As risk models are often probabilistic in nature, the loss function to examine is an *expected* predictive loss. Thus, the assessment of closeness are usually defined using the language of expectation:

$$\mathbb{E}\left(\mathcal{L}\left(S_p\left(\vec{X}_p;\hat{\theta}_p\right),Y\right)\right),$$

where the expectation may be taken over many sources of uncertainties. A common choice of the loss function is the squared error loss, $\mathcal{L}(a,b)=(a-b)^2$. The squared differences for each prediction are summed and divided by the total number of predictions (or the total number of individuals about whom a prediction is made). This computation gives the average squared error loss. The lower the loss, the more accurate the model. The square error loss is known as the Brier score. It is a strictly proper scoring rule, and is the loss function employed in this paper. There are other choices of loss functions that may be particularly indicative of model performance in different contexts, such as the Area Under Curve (AUC) or the Matthew correlation coefficient.

[30] Recall that true attributes are those that, as a matter of fact, bring about the outcome of interest in the generative (oracle) model.

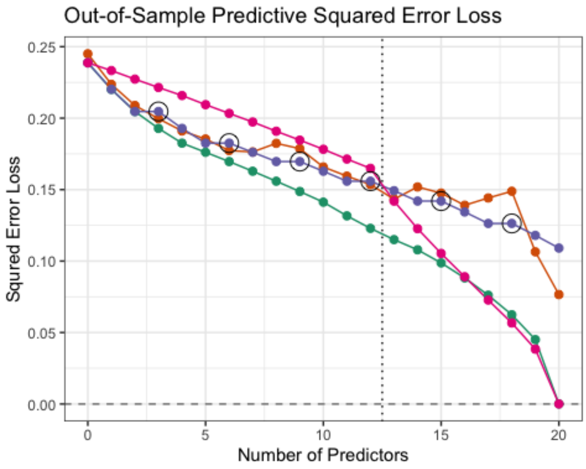

Fig. 3 Out-of-sample predictive squared error loss as a function of the number of predictors. Models represented by the green curve use a large training set ($n = 10^5$) with the correct predictor variables, for which the group-dependent and group-independent attributes are entered in a mixed order. Orange: same as green but a small training set ($n = 100$) is used. Magenta: same as green but the group-independent attributes (12 in total) entered before the group-dependent attributes (8 in total). Purple: same as green but six of the predictors (circled) are mis-specified. The only empirical model that reaches perfect accuracy is represented by the last point on both the green and the magenta curves.

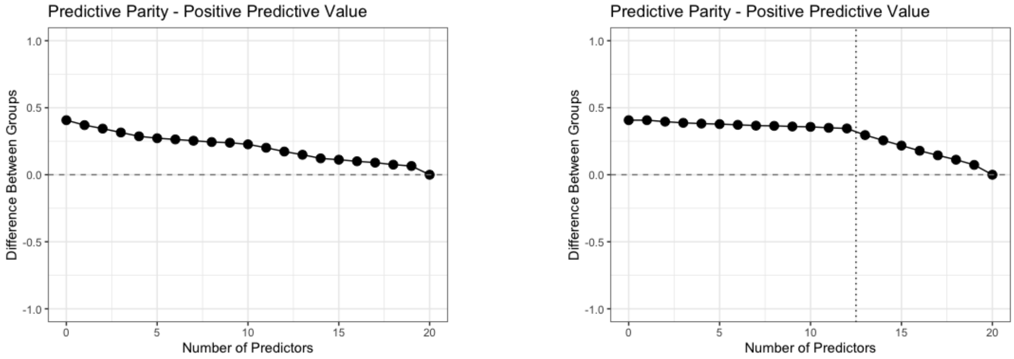

Fig. 4 The predictive parity of the empirical risk models as measured in terms of the positive predictive value (PPV). Left panel: models are trained according to the standard simulation setup (green curve of Figure 3) with a large training set and correct predictors entering into the model in a mixed fashion. Right panel: the group-dependent (but correct) features are excluded for the first 13 models.

informational richness means better accuracy as well as better fairness if the latter is understood as predictive parity. This monotonic trend exists assuming several idealizations as part of the simulation, for example, that all predictors are true predictors of the outcome of interest.

The outlier here is classification parity, another popular measure of algorithmic fairness. So the third trend that emerges from the simulation is that classification

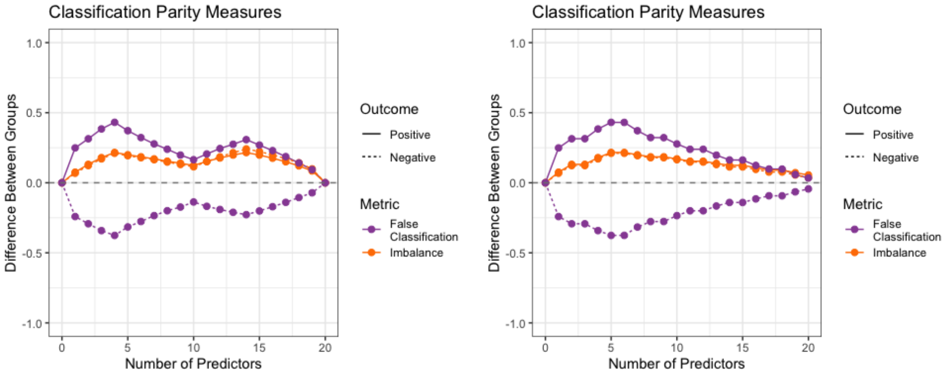

Fig. 5 Classification parity as measured in terms of false positive rate, false negative rate and imbalance for the sequence of empirical risk models using a large training set and the correct predictors (left; corresponding to the green curve of Figure 3) versus the partially misspecified predictors (right; corresponding to the purple curve). Classification parity is perfect only in two extreme scenarios: either the empirical model has perfect accuracy (rightmost point in left figure) or the worst accuracy (baseline; leftmost point in both left and right figures).

parity behaves erratically. Recall that classification parity, unlike predictive parity, requires that the rate at which people are correctly classified as 'positive' or 'negative' be the same across groups. Classification parity does not monotonically improve as a result of better accuracy and richer information. It is achieved when the risk model relies on no predictors at all and consequently when the model's accuracy is at its worst. Tossing a fair coin to decide how to classify individuals would always deliver classification parity. Classification parity can also be achieved with full information and perfect accuracy. As we learn from our simulation, however, anything between full and null information fails to deliver classification parity. But, unlike predictive parity, adding more predictors may reduce classification parity in some cases and improve it in other cases (Figure 5).

These trends show that achieving classification parity conflicts with the objective of achieving an accurate prediction of the outcome of interest. The same conclusion does not hold for the other metrics of fairness performance, in particular predictive parity. This fact is a good reason to be wary of classification parity as a measure of algorithmic fairness. We will spell out an argument to this effect in the next section. But before doing that, it is paramount to situate performance criteria of fairness in relation to what we have been calling conscientiousness and informational richness.

Conscientiousness is a function of all aspects that are under the control of those in charge of constructing the risk model. One salient aspect of conscientiousness is the number of true predictors used in the risk model, what we have called informational richness. Another aspect is the representativeness of the data. A third aspect still is the selection of the appropriate model along with its parameters. For simplicity, we focus on informational richness, keeping in mind that this is just one dimension of conscientiousness. This perspective affords us a new reading of what is going on in some of the counterexamples in the philosophical literature against performance criteria of fairness.

Recall Hedden's example in which people are assigned coins with different biases (the objective risk) and the predictive algorithm assigns risk scores to each person based on these biases. Hedden shows that playing with differences across two groups of interest—differences in the distribution of objective risks or differences in prevalence—is enough to ensure that many performance measures are violated (see earlier discussion in Section 2). But this observation also works in the other direction: playing with differences across groups can ensure that certain performance measures are satisfied. For suppose a credit algorithm violates equality of false positive classification across two groups, say, group $G = 1$ has a higher rate of false positives than $G = 0$. To correct this disparity, a bank decides that the pool of applicants from $G = 1$ should include more people who are credit-worthy and who can be easily classified as such, thanks to characteristics such as stable income and timely credit card payments. Because of this, the false positive rate for $G = 1$ will go down and be equalized to that of group $G = 0$. The same manipulation can be carried out to ensure compliance or violation of other performance criteria of fairness.

So, violating as well as satisfying performance criteria of algorithmic fairness sometimes comes relatively *easily* provided one artificially changes in the right way the composition of the two groups being compared. When these artificial changes are made, fairness criteria of performance appear to be divorced from our intuitions about fairness. After all, *ad hoc* manipulations should have no effect on the fairness of the algorithmic predictions. This is so—we hold—precisely because artificially changing the composition of the groups does not require an improvement in conscientiousness. Sometimes compliance with performance criteria of fairness is even obtained by openly disregarding conscientiousness. Consider calibration, a form of predictive parity. Suppose the prevalence rate of the outcome of interest in group $G = 1$ is 70% and only 40% in $G = 0$. A predictive algorithm could assign a risk score of .7 to every person in $G = 1$ and .4 to every person in $G = 0$. The algorithm would be calibrated, since the fraction of people who are actually positive in each group would correspond to the score assigned to them.[31] But this calibration could hardly be indicative of a fair algorithm.[32] It could not be indicative of fairness—we hold—precisely because the risk score was estimated in a coarse manner, most likely giving up information available, an open violation of conscientiousness.

Should we then give up on performance criteria of fairness and focus exclusively of conscientiousness? This would be too quick. As seen above, accuracy, conscientiousness and predictive parity go hand in hand. So an improvement along one dimension can be indicative of an improvement in another dimension. And sometimes compliance with performance criteria may be more easily verifiable than a multifaceted idea such as conscientiousness or informational richness. Still, if it is clear that no improvement in conscientiousness or accuracy has taken place, the improvement in predictive parity must be the result of a form of manipulation that has little to do with the fairness of the algorithm.

The case of classification parity is different, however. For compliance with this performance criteria may sometimes even require one to sacrifice conscientiousness

[31] On the definition of calibration, see footnote 7.

[32] A similar example was given by Corbett-Davies and Goel (2018) in a seminal paper on algorithmic fairness.

and base one's prediction on a smaller set of predictors. Unlike predictive parity, classification parity comes into open conflict with accuracy and conscientiousness. This is a strong reason to dispense with classification parity altogether. This is the topic of the next section.

5 Against Classification Parity

We consider two proposals for defending classification parity, and we find both of them unsatisfactory. First, classification parity might be an appealing criterion of algorithmic fairness insofar as it tracks the comparative probability of misclassification to which people from different groups are subject. On this interpretation, classification parity is bottomed in the expectation that people from different groups should have equal prospects of misclassification, where such prospect is understood as the probability of misclassification. That people in minority groups are more likely to be misclassified because of their group membership seems unfair, especially when a higher probability of misclassification translates into a higher risk of harm, such as being erroneously placed in preventative detention.

But deviations from classification parity do not necessarily entail an uneven allocation of the prospects of misclassification across individuals belonging to different groups of interest. Here, what we mean by 'group' is specifically the result of applying a protected category such as race or gender. Take an individual in a minority group and compare this individual with another who possesses the same predictive features but belongs to a non-minority group. These two individuals will be treated the same, either correctly classified or not. This is just how algorithms work: people with the same predictive features are treated the same. So one individual would not be more likely to be misclassified than the other, even though they belong to different groups.[33]

In the aggregate, as our simulation shows, individuals in one group will be *more or less often* misclassified than individuals in another group, assuming the base rates of the outcome of interest differ between the two groups. Thus, one might argue that, because of these differences in the frequency of misclassification, the individuals in one group are more likely to be misclassified than the individuals in another group. But, crucially, these judgments of probability hold for *average* individuals who are described not by the full set of features available to characterize them. The assessment of the probability of misclassification should instead be relative to the most specific description available to the algorithm. This description will not be uniquely individualizing, but will likely include both group-dependent and group-independent predictive features. In this case, individuals who belong to different groups and are otherwise similar under the most specific description cannot be subject to uneven prospects of misclassification.

[33] Long (2021) makes the point that group differences in false positive rates do not track group differences in the risks (prospects) of error. To make this point, Long relies on a hypothetical case (see footnote 23) in which group membership is causally irrelevant to whatever features are used by the algorithmic to make its predictions. The argument here does not make this assumption. In our simulation study (see Section 3), group membership is causally implicated in bringing about some of the predictors used by the predictive algorithm.

In response, some might object that this argument looks at classification parity too narrowly, in isolation from the larger trends in society. Violations of classification parity can very well be an indication of antecedent group-based systemic disadvantage—that people who are worse-off are deprived of opportunities even when they should not be. Suppose, for the sake of illustration, that black people described at the most individualized level from the perspective of the predictive algorithm—say, viewed in light of their income, health, education—are not subject to a higher probability of false loan rejection compared to similarly situated white people. And yet, suppose black people on average—not described at most individualized level—have a higher probability of false loan rejection. A plausible explanation for this violation of classification parity is that black people at an antecedent stage of life had lower chances to have access to adequate income, education, health care. These lower chances are reflected into higher rates, on average, of false loan rejections against black people compared to white people.

The above illustration suggests that, far from being morally irrelevant, violations of classification parity reflect larger disparities in societies. But how general is this phenomenon and does it follow a predictable pattern? Some have claimed that the group with higher prevalence—which, in some contexts, is the disadvantaged group—will suffer higher rates of false positives and lower rates of false negatives, which may entail higher rates of police stops, jail times, and mortgage rejections.[34] If the group subject to a higher rate of false positives and a lower rate of false negatives is the disadvantaged group, violations of classification parity would follow a predictable pattern that mirrors larger social inequalities.

Our simulation, however, shows a more complex picture. Depending on the type of predictors used by the algorithm—group-dependent or group independent features—the group with higher prevalence may be subject to a lower rate of false positives and a higher rate of false negatives, as is apparent by comparing Figure 5 and Figure 6. So classification parity need not systematically mirror larger disparities in society. Perhaps the argument here is that classification parity matters because its violation tends to entrench historical injustices in societies along racial or gender lines. But note that classification parity is the conjunction of two requirements: equal false positive and equal false negative rates. Classification parity would be violated whether a minority group is subject to a higher or lower rate of loan false rejections. Arguably, only a higher rate would entrench injustices in access to credit to the detriment of a minority group. A lower rate of loan false rejections could actually disrupt historical injustices. In any event, a more careful analysis of the relationship between historical injustices and violations of classification parity is needed.

So, to summarize, we contend that classification parity, as a criterion of algorithmic fairness, should be given limited weight. Our claim rests on three observations. First, the simulation shows that pursuing classification parity, unlike predictive parity, is a somewhat erratic goal, as it may conflict with improving accuracy and informational

[34] For example, consider Northpointe's answer in (Dieterich et al., 2016) to ProPublica's accusation in (Angwin et al., 2016) that COMPAS is racially biased. COMPAS is an algorithm used in several jurisdictions in the United States to make predictions about recidivism. Northpointe alleged that, since the prevalence of criminality among black people is higher, false positives will also be higher and false negatives will be lower. Long (2021) makes the same claim with the qualification that it holds if (a) the algorithm meets predictive parity and (b) it applies the same decision threshold for different groups.

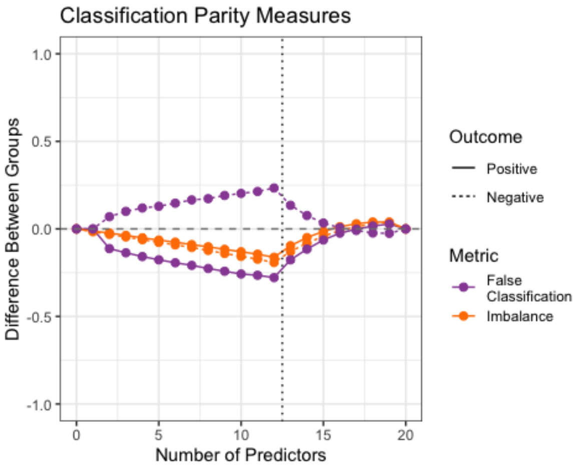

Fig. 6 If all true predictors that are correlated with group membership are excluded from the empirical model (the first 13 points on each curve), the group with a higher prevalence (here $G = 1$) receives a lower false positive rate and a higher false negative rate. This is a departure from classification parity, albeit in the opposite direction, compared to the models in Figure 5 for which all predictors, true or otherwise, enter in a mixed order regardless of their dependence on group.

richness. In addition, its erratic behaviour is paired with the fact that classification parity does not track any tangible disparity in the prospects of misclassification across individuals in different groups. Finally, as our simulation shows, violations of classification parity need not always go to the detriment of the disadvantaged group. They may entrench as well as disrupt historical injustices. So, the balance of reasons weighs against classification parity as an intuitively appealing criterion of algorithmic fairness.

We conclude this section by pointing out that—inevitably—predictive algorithms will subject people to uneven prospects of mistaken classification, but not in the way that violations of classification parity might suggest. This problem is pervasive, but we think it is best addressed by reasserting the centrality of informational richness.

Suppose we compare two groups of individuals: one group comprises people who possess all features that are positively correlated with the outcome and the other group comprises individuals who possess all features that are negatively correlated. Suppose we select only people from these two groups that will not bring about the outcome of interest, such as defaulting on a loan or committing a crime. Still, because of the different features they possess, the people in one group will be incorrectly labeled as 'positive' and the others correctly labelled as 'negative'. Thus, the people in the two groups are subject to uneven prospects of misclassification.[35] This disparity raises a

[35] For an argument about uneven prospects of mistaken convictions in criminal trial and its implications for fairness, see Di Bello and O'Neil (2020). The argument (roughly) is this. Suppose, for example, there is profile evidence that shows that low socioeconomic status is positively correlated with the crime of drug trafficking. If you are on trial for drug trafficking and are of low socioeconomic status, should this profile evidence be introduced as evidence against you? It would seem unfair to present this evidence against you. One way to make sense of this unfairness is to realize the following fact: if you were innocent, you would be mistakenly convicted with a greater probability than those of higher socioeconomic status against whom the same profile evidence could not be used as incriminating. After all, if the profile evidence were added to other evidence available against you at trial, this addition may tip the balance of the evidence in favor of a conviction. So, in this context, if you were an innocent facing trial, you would be more likely to be mistakenly

fairness concern, but one that has little or nothing to do with violations of classification parity. The disparity at issue concerns individuals viewed at the most specific level of description available. The disparity cuts across protected categories and may occur within the same protected group.

The problem for predictive algorithms just identified stems from the reliance on correlations between a select set of features and an outcome. Some individuals who are factually negative will be indistinguishable—from the point of view of the algorithm—from some factually positive individuals and thus the algorithm will classify them the same. This problem can be avoided by increasing the ability of the predictive algorithm or risk model to distinguish between otherwise indistinguishable individuals. A more fine-grained set of predictors could tame the unavoidable fact that certain individuals, described in the most specific way, are subject to uneven prospects of misclassification. But such refinement might not always be possible in practice, as we discuss in the next section.

6 Cautionary warnings

This paper focused on the centrality of informational richness and more generally conscientiousness. Risk models make predictions about individuals on the basis of a set of predictors. The greater informational richness, the greater the accuracy of the risk model, the greater its predictive fairness. So, informational richness is the engine that drives improvements in the performance of a predictive model, in terms of model accuracy and fairness. The results from the simulation make the centrality of informational richness vivid. The outlier here is classification parity whose performance is erratic.

But the centrality of informational richness should be qualified, and this final section adds some cautionary warnings. The first warning is that there are two sources of uncertainty that predictive algorithms or risk models should attempt to contain:

(a) A risk model will not be entirely correct whenever not all true or relevant predictors are included in it or some of the predictors included in the model are not true or relevant predictors. Call this *informational uncertainty.* This uncertainty is progressively eliminated as the algorithm knows more and more aspects of what should be known.

(b) Another source of error for the predictions made by the risk model is *data uncertainty.* Even if the data is unbiased and representative, it could still be too small to be used to construct a risk model that makes reliable predictions.[36]

These two sources of uncertainty form a trade-off. It is a good idea to base predictions on as many true predictors as possible. This will reduce informational uncertainty. Gathering more information is in principle always possible, but statistical inference faces an inescapable limit when it is applied to the behaviour of individuals. Even

convicted. The analogy with algorithmic predictions should be clear. They rest on a very sophisticated form of profile evidence that involves multiple correlations between certain features and an outcome of interest.

[36] Another source of uncertainty is *modeling uncertainty.* The model could be mis-specified, in the sense that it does not capture the structure of true data generating process. When this happens, even in absence of informational or data uncertainty, the risk model will fail to approximate perfect accuracy. See Figure 3 (orange line).

though the goal is to determine the risk *this* individual will do this or that, the risk can only be statistically estimated by making comparisons and generalizations from features that are also shared by other individuals. Since any richer set of predictors will be instantiated by a smaller class of people, the reliability of these generalizations will also inevitably decline. So the more predictors, the smaller the sample size, the greater the data uncertainty.[37]

What are the implications of this trade-off between informational and data uncertainty? There are some aspects of human decision-making that no predictive algorithm can capture, not because data about them cannot in principle be collected, but rather, because there would not be enough data to make reliable predictions. Even our best picture could still fail to capture the world in all its complexity. It is possible, for all we know, that two people share the exact same attributes, and yet go on to bring about different outcomes.

To capture this practical and conceptual limitation of predictive algorithms, we amended the data generating process in the simulation. We have assumed thus far that the data generating mechanism is deterministic—that a specific set of features uniquely determines the outcome of interest and that the set of features, in its entirety, is knowable in principle. This setup assumes that predictive algorithms are capable of capturing the relationship between predictive features and outcome in the finest detail possible. But this assumption breaks down because of the trade-off between informational and data uncertainty. So, we changed the data generating mechanism by stipulating that the relationship between predictive attributes and outcome is not deterministic, but governed by the flip of a weighted coin, where the probability of the outcome is a function of the attributes and the coefficients. Specifically, we revised the generative model in equation (1) from Section 3 into the following:

$$S_\infty = \text{Probit}^{-1}\left(\beta_0^* + \beta_1^* X_1 + \cdots + \beta_{20}^* X_{20}\right), \quad Y \sim \mathit{Bernoulli}(S_\infty). \tag{2}$$

According to the revised model, given two individuals exactly identical as far as what is knowable about them, the outcome of interest could still be different.[38]

We do not intend to suggest here that human decision-making is random or indeterministic. Rather, the stochastic data generating process makes vivid the realization that predictive algorithms are not capable—not even in principle—of capturing the relationship between predictive features and the outcome of interest in the finest detail possible. This incapability has a serious detrimental effect on algorithmic performance in terms of both accuracy and fairness. The simulation in its amended version shows that, once we give up the determinism assumption in how the data are generated, but maintain the same, correctly specified empirical risk model, every performance indicator on the side of fairness or accuracy deteriorates (Figure 7 and 8).

So, on the assumption of a stochastic relationship between predictive attributes and outcome, even predictive algorithms that perfectly approximate the objective risk will

[37] This trade-off between informational and data uncertainty is also known as the bias-variance trade-off (Li and Meng, 2021).

[38] Incidentally, Hedden's scenario in which each individual is associated with a coin whose bias represents the objective chance for the individual of bringing about the outcome assumes a stochastic relationship between predictive attributes and outcome.

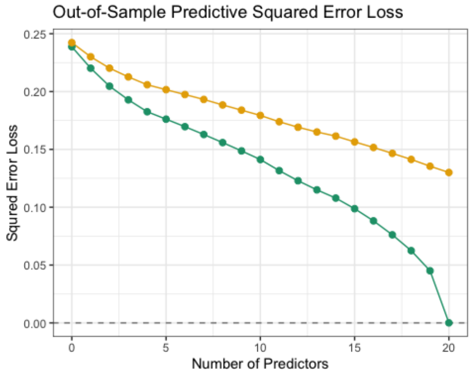

Fig. 7 Predictive square error loss (yellow curve) of the sequence of empirical risk models fitted with the correct predictors, but with formula (2) as the data generative model. Here, the idealized risk score is a fractional number between [0, 1] and is no longer deterministically associated with the individual's outcome Y. Even the best model (i.e. the model with all 20 true predictors) cannot achieve perfect accuracy. The green trajectory is the same as that in Figure 3 and represents out of sample squared error for the deterministic scenario.

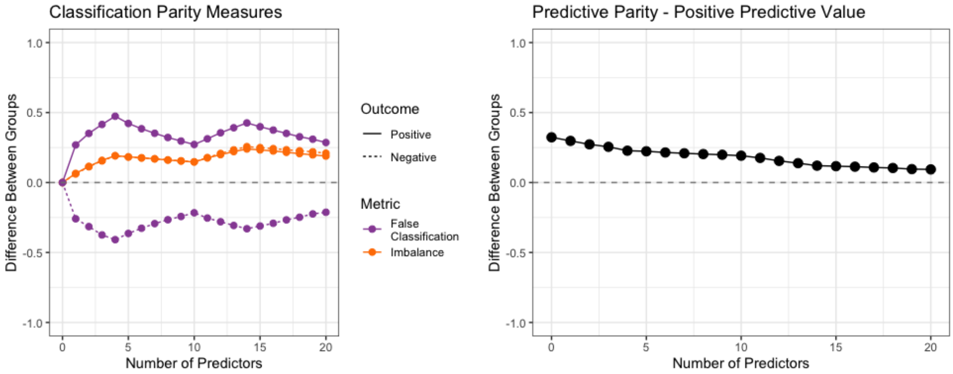

Fig. 8 Classification parity (left) and predictive parity (right) for the sequence of empirical risk models fitted with the correct predictors with formula (2) as the generative model.

inevitably exhibit disparities in performance across groups. Call this *baseline performance disparity*. We can draw a moral from this fact. Instead of insisting on compliance with fairness performance measures in the absolute, aiming to progressively approximate the baseline performance disparity might be a more meaningful objective. This can be done following informational richness. After all, it is still possible to progressively improve performance by adding more predictors, even though perfect parity in performance is never reached no matter how many true predictors are included.

The second cautionary warning is that information richness should not be simply equated with the amount of data, or types of data, that are available. If the risk model itself is not thoughtfully specified or carefully estimated, there is no guarantee that it would become more accurate with increasing amount of input data, even if the data are indeed informative. While a loss of accuracy resulting from more data may be paradoxical, models that suffer from it may be employed in practice, such as those that lack *self-efficiency.*[39] Without self-efficiency, adding more data to the risk model may not improve the performance in terms of accuracy and fairness. Our simulation utilizes well-studied statistical models and estimation procedures that do no suffer from a lack of self-efficiency, but such a quality should not be taken for granted. In addition, we note that with respect to a given class of risk models, all additional predictors may not be true or relevant predictors. Some may be uncorrelated with the outcome of interest or even misleading. In our limited simulation analysis, adding misspecified predictors has zero net effect on predictive accuracy (Figure 3). Neither accuracy improves nor does it worsen. One may be tempted to conclude that all else being equal, it is better to add additional predictors to the risk model, since in the worst case scenario they would simply have a zero net effect for predictive accuracy. But, everything else is *not* equal, as the additional costs of gathering larger datasets to sustain the same model quality, in terms of time, money or heightened intrusion into people's privacy, are not negligible. The modeler must balance the trade-off between the costs of relying on additional predictors and their added value for accuracy and fairness.

These two cautionary warnings are strongly related. Conceptually, there is a limit to the number of predictors the risk model can rely on because datasets will be inevitably smaller the more predictors are used. Practically, the size of datasets will be constrained by costs, in terms of money, time and privacy. These cautionary remarks, however, should not detract from the key message of our discussion: informational richness should take more prominence in the literature on algorithmic fairness.

7 Conclusion

The argument in this paper mostly centered on how the informational richness of the predictors used by risk models affect performance criteria of accuracy and fairness. As seen in Section 3 and 4, the more true predictors are used, the better the accuracy and fairness of the algorithm, leaving aside classification parity for the reasons we give in Section 5. We conclude by sketching how informational richness can help to shed light on other notions of algorithmic fairness.

Besides performance criteria of fairness, we think that informational richness is also relevant to *attitudinal* criteria of algorithmic fairness, for example, the requirement that the same risk threshold be applied to different individuals belonging to different groups. Algorithms strictly speaking do not have attitudes, but the humans who design them and put them to use certainly do. Consider an algorithm who makes predictions about the risk of loan default. If people in some groups needed a lower risk threshold to qualify for a loan than people in another group, this would signal a prima facie difference in attitudes, say, that the costs of erroneous decisions were weighed differently for people in different groups.[40] But this fairness criterion of 'same threshold' risks

[39] Self-efficiency requires the model's estimate to be more accurate when computed using the complete dataset. A model is not self-efficient if its estimate achieves a smaller mean squared error when applied to a subset of data selected from the complete data (Meng, 1994). Xie and Meng (2017) discuss cases in which the lack of self-efficiency arises in the context of multi-phase statistical inference.

[40] Same threshold is often taken for granted as a criterion of fairness. For an examination of this criterion of algorithmic fairness, see Johnson King and Babic (ms). On the other hand, Aziz Huq Huq (2019) argues that, in some cases, fairness requires that same threshold be violated. Huq points out that people in minority

being an empty shell if it is not accompanied by another attitudinal criterion, what we might call *equal conscientiousness.* As a first pass, think of equal conscientiousness as the requirement that, across individuals belonging to different groups, the predictive algorithm relies on an equally rich set of predictive features. If different individuals were assessed by the algorithm with uneven conscientiousness—that is, using richer or poorer sets of predictors—this would signal a difference in attitudes toward them, as though some people were deserving a more careful risk assessment than others. This difference in attitudes would exist even if—nominally—the same risk threshold were still applied across groups. So the requirement of applying the same threshold is best accompanied by the requirement of equal informational richness.

But there is a further complication here. It might be appropriate to rely on a larger set of predictors for one group versus another insofar as these different sets of predictors perform equally accurately across the two groups. So equal conscientiousness might actually require reliance on different set of predictors. Recall that our simulation shows the following: the more true predictors, the more accurate the risk model. But it also shows that the speed of this monotonic improvement is not the same across groups. The same level of accuracy is reached via a smaller set of predictors for one group compared to another (Figure 3). Given these differences in performance, equal conscientiousness might require that the predictive algorithm rely on a varying number of predictors depending on which group the individual who is the target of the prediction belongs to.

These remarks suggest that attitudinal criteria of fairness—such as equal threshold—should not be understood independently of performance criteria of accuracy and fairness. And—we hold—informational richness and conscientiousness can help to draw the relevant connections. But a more general examination of the relationship between performance criteria and attitudinal criteria of algorithmic fairness is left for another time.

Acknowledgments. For comments on earlier drafts, we would like to thank Brian Hedden, Katie Steele, Colin O'Neil, Jan-Willem Romeijn, Boris Babic, Zoë Johnson King, Fabrizio Cariani, three anonymous reviewers of this journal, audiences at the University of Lausanne and the Australian National University. In addition, we would like to thank Claire Benn, Todd Karhur, Seth Lazar and Pamela Robinson for organizing a special issue on Normative Theory and Artificial Intelligence. Finally, we would like to thank Deborah Mayo for organizing a summer school on the philosophy of statistics at Virginia Tech in the summer of 2019. She made the collaboration between a statistician and a philosopher possible.

References

Aaronson, D., J. Faber, D. Hartley, B. Mazumder, and P. Sharkey. 2021. The long-run effects of the 1930s holc "redlining" maps on place-based measures of economic opportunity and socioeconomic success. *Regional Science and Urban Economics* 86: 103622 .

groups may suffer greater harm as a result of a mistaken algorithmic classification, and when this is the case, the decision threshold should be more stringent for them.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing .

Barabas, C., J. Bowers, J. Buolamwini, R. Benjamin, M. Broussard, S. Constanza-Chock, K. Crawford, C. Doyle, B.E. Harcourt, B. Hopkins, M. Minow, R. Ochigame, T. Priyadarshi, B. Schneier, J. Selbin, K. Dinakar, T. Gebru, S. Helreich, J. Ito, C. O'Neil, H. Paxson, R. Richardson, J. Schultz, and V.M. Southerland. 2019. Technical flaws of pretrial risk assessment raise grave concerns.

Barocas, S. and A.D. Selbst. 2016. Big data's disparate impact. *California Law Review 104*(3): 671–732 .

Berk, R., H.H. an d Shahin Jabbari, M. Kearns, and A. Roth. 2021. Fairness in criminal justice risk assessment: The state of the art. *Sociological Methods and Research 50*(1): 3–44 .

Binns, R. 2020. On the apparent conflict between individual and group fairness. FAT* '20, pp. 514–524. Association for Computing Machinery.

Borsboom, D., J.W. Romeijn, and J.M. Wicherts. 2008. Measurement invariance versus selection invariance:is fair selection possible? *Psychological Methods 13*(2): 75–98. https://doi.org/https://doi.org/10.1037/1082-989X.13.2.75 .

Cai, W., J. Gaebler, N. Garg, and S. Goel 2020. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.*

Castro, C. 2019. What's wrong with machine bias. *Ergo 6*(15): 405–426 .

Chen, I., F.D. Johansson, and D. Sontag. 2018. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*: 3543–3554 .

Chen, J., S.S. Rathore, M.J. Radford, Y. Wang, and H.M. Krumholz. 2001. Racial differences in the use of cardiac catheterization after acute myocardial infarction. *New England Journal of Medicine 344*(19): 1443–1449 .

Chiappa, S. and T.P.S. Gillam. 2018. Path-specific counterfactual fairness.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* https://doi.org/10.1089/big.2016.0047 .

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

Dehon, E., N. Weiss, J. Jones, W. Faulconer, E. Hinton, and S. Sterling. 2017. A systematic review of the impact of physician implicit racial bias on clinical decision making. *Academic Emergency Medicine 24*(8): 895–904 .

Di Bello, M. and C. O'Neil. 2020. Profile evidence, fairness, and the risks of mistaken convictions. *Ethics 130*(2): 147–178 .

Dieterich, W., C. Mendoza, and T. Brennan. 2016. Compas risk scales: Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county .

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel 2012. Fairness through awareness. In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226.

Green, B. 2022. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology 35*(90): 1–32 .

Gross, S.R., M. Possley, K. Otterbourg, K. Stephens, J.W. Paredes, and B. O'Brien. 2022. Race and wrongful convictions in the united states 2022.

Hedden, B. 2021. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs 49*(2): 209–231. https://doi.org/10.1111/papa.12189 .

Heidari, H., M. Loi, K.P. Gummadi, and A. Krause 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. FAT* '19, New York, NY, USA, pp. 181–190. Association for Computing Machinery.

Hellman, D. 2021. Big data and compounding injustice. *Virginia Public Law and Legal Theory Research Paper No. 2021-27* .

Huq, A.Z. 2019. Racial equity in algorithmic criminal justice. *Duke Law Journal 68*(6): 1043–1134 .

Johnson King, Z. and B. Babic. ms. Algorithmic fairness and resentment .

Jorgensen, R. 2022. Algorithms and the individual in criminal law. *Canadian Journal of Philosophy 52*(1): 61–77 .

Kearns, M. and A. Roth. 2019. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design Illustrated Edition.* Oxford University Press.

Kleinberg, J., S. Mullainathan, and M. Raghavan 2017. Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitrou (Ed.), *8th Innovations in Theoretical Computer Science Conference*, Volume 43, pp. 43:1–43:23.

Kusner, M.J., J.R. Loftus, C. Russell, and R. Silva. 2018. Counterfactual fairness.

Ladd, H.F. 1998. Evidence on discrimination in mortgage lending. *The Journal of Economic Perspectives 12*(2): 41–62 .

Lazar, S. and J. Stone. ms. On the site of predictive justice .

Lee, M.S.A., J. Singh, and L. Floridi. 2021. Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. *AI Ethics* 1: 529–544 .

Li, X. and X.L. Meng. 2021. A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association 116*(533): 353–367 .

Lippert-Rasmussen, K. 2011. We are all different: Statistical discrimination and the right to be treated as an individual. *Journal of Ethics 15*(1-2): 47–59 .

Long, R. 2021. Fairness in machine learning: Against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy 19*(1): 49–78 .

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, , and A.H. Byers. 2011. Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute* .

McKinlay, J.B. 1996. Some contributions from the social system to gender inequalities in heart disease. *Journal of Health and Social Behavior 37*(1): 1–26 .

Meng, X.L. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*: 538–558 .

Menon, A.K. and R.C. Williamson 2018. The cost of fairness in binary classification. In S. A. Friedler and C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118. PMLR.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application 8*(1): 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902 .

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464): 447–453. https://doi.org/DOI:10.1126/science.aax2342 .

Petersen, L.A., S.M. Wright, E.D. Peterson, and J. Daley. 2002. Impact of race on cardiac care and outcomes in veterans with acute myocardial infarction. *Medical Care 40*(1): I–86–I–96 .

Powers, M. and R. Faden. 2019. *Structural Injustice: Power,Advantage, and Human Rights*. Oxford University Press.

Raghupathi, W. and V. Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems 2*(3): 1–10 .

Rehavi, M.M. and S.B. Starr. 2014. Racial disparity in federal criminal sentences. *Journal of Political Economy 122*(6): 1320–1354 .

Reich, C.L. and S. Vijaykumar 2021. A possibility in algorithmic fairness: Can calibration and equal error rates be reconciled? In K. Ligett and S. Gupta (Eds.), *2nd Symposium on Foundations of Responsible Computing, FORC 2021*, Volume 192 of *LIPIcs*, pp. 4:1–4:21.

Schulman, K.A., J.A. Berlin, W. Harless, J.F. Kerner, S. Sistrunk, B.J. Gersh, R. Dubé, C.K. Taleghani, J.E. Burke, S. Williams, J.M. Eisenberg, W. Ayers, and J.J. Escarce. 1999. The effect of race and sex on physicians' recommendations for cardiac catheterization. *New England Journal of Medicine 340*(8): 618–626 .

Sharifi-Malvajerdi, S., M. Kearns, and A. Roth 2019. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing System*, Volume 32, pp. 8242–8251.

Slobogin, C. 2021. *Just Algorithms: Using Science to Reduce Incarceration and Inform a Jurisprudence of Risk*. Cambridge University Press.

Suresh, H. and J. Guttag 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21. Association for Computing Machinery.

Vigano', E., C. Hertweck, C. Heitz, and M. Loi 2022. People are not coins. morally distinct types of predictions necessitate different fairness constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2293–2301.

Xie, X. and X.L. Meng. 2017. Dissecting multiple imputation from a multi-phase inference perspective: what happens when god's, imputer's and analyst's models are uncongenial? *Statistica Sinica*: 1485–1545 .

Young, I.M. 2003. Political responsibility and structural injustice. *The Lindley Lecture, University of Kansas* .

Zimmermann, A. and C. Lee-Stronach. 2022. Proceed with caution. *Canadian Journal of Philosophy 52*(1): 6 – 25 .