



DOI:10.1145/3587930

**Standards for fair decision making could help us develop algorithms that comport with our consensus views; however, algorithmic fairness has its limits.**

BY MANISH RAGHAVAN

# What Should We Do when Our Ideas of Fairness Conflict?

THERE IS GROWING interest in using algorithms to make fair decisions. However, recent results prove that different notions of fairness cannot be simultaneously satisfied.<sup>10,25</sup> In this article, we explain and explore the consequences of these impossibility results. The literature contains a variety of responses. Some propose technical relaxations or reformulations that admit solutions, others argue to prioritize some measures of fairness at the expense of others, and still others hold that formal impossibility results should

force us to re-examine the broader contexts within which algorithms are deployed. Here, we survey these responses and discuss their implications for the use of algorithms in decision making.

We are constantly faced with decisions in our daily lives. Some appear fairly inconsequential: an ad shown before the next video you watch or the sequence of posts on your social media feed. Others can change our lives—for example, whether we get a certain job or are approved for a loan. Algorithms play a growing role in these types of decisions. In response, a nascent field has formed, bridging disciplines such as computer science, economics, sociology, and legal studies in an effort to understand the impact of algorithmic decision making on society.<sup>34</sup>

One key area within this field considers fair decision making. When algorithms are used to make or assist with consequential decisions, how do we ensure that they do so fairly? This question is particularly salient when it comes to machine learning and other data-driven tools, where we might expect algorithms trained on data produced by humans to inherit the same biased and discriminatory behavior that humans exhibit. Researchers and practitioners have begun developing tools to address concerns over these behaviors, often using phrases like “algorithmic fairness” or “fairness in machine learning” to describe their efforts.

A key challenge to this work is the ambiguity behind the word “fair.” It has no precise meaning, nor does soci-

## >> key insights

- **Attempts to mathematically formalize fairness reveal seemingly reasonable definitions of fairness can be mutually incompatible.**
- **In some contexts, mathematical or philosophical arguments can be made to relax or prioritize particular constraints.**
- **In others, recognizing limitations inherent to fairness measures offers the opportunity to explicitly reconsider the broader contexts in which predictive systems are deployed.**





# fairness

\ 'fa(ə)r-nəs \

(noun)

1. the quality or state of being reasonable, right, and just
2. the ability to make judgments free from discrimination or dishonesty

ety necessarily agree on what it entails. For example, in higher education, some consider affirmative action to be the only way to ensure fairness, while others believe affirmative action to be inherently unfair. Still, despite the lack of clarity on exactly what constitutes fair decision making, we might hope to impose some basic standards on the algorithms we build to ensure they respect intuitive tenets of fairness. If we could at least agree on some minimal standards for fair decision making, then we could develop algorithms that comport with our consensus views, even if they do not completely resolve our normative disagreements.

Defining basic standards or guardrails by which algorithms should abide has become a common approach in the literature—for example Dwork et al.<sup>16</sup> and Hardt et al.<sup>21</sup> Much of the work in this vein seeks to formally specify a measurement that captures some notion of fairness and develop algorithms that are fair according to this measurement.

For example, early work by Dwork et al.<sup>16</sup> considered the principle that similar individuals ought to be treated similarly and operationalized it through an algorithmic framework. Others have developed techniques to mitigate disparities in predictive performance measured across demographic groups.<sup>21</sup> We commonly refer to this idea, comparing measures of performance disaggregated by demographic groups, as “group fairness,” and it will be the primary focus of this article.

Metrics that aim to capture elements of fair decision making are necessarily incomplete. Normative beliefs about fairness are far too complex and nuanced to encode in coarse mathematical constraints. Moreover, they operate under incomplete information: Data fails to encode much of the important context—such as mitigating circumstances, access to opportunity, and intent—that guides our intuitions about fairness. Many proposed measures in the literature require ag-

gregating over broad, heterogeneous swathes of the population, potentially ignoring harms experienced at an individual level. Perhaps more worryingly, as Powles and Nissenbaum<sup>33</sup> argue, a focus on technical formulations of fairness may obscure broader ethical and social concerns; a slightly fairer way of engaging in an unethical activity is still unethical. As we will see, many scholars contend that the types of reforms suggested by algorithmic fairness are insufficient to bring about substantive change, which requires a reconsideration of the social systems within which algorithms operate.<sup>13,18</sup>

And yet, we still use algorithms in impactful decisions, and in such cases, we have an interest in detecting and preventing normatively undesirable behavior, even if we cannot hope to resolve ethical challenges through technical means alone. Abebe et al.<sup>1</sup> describe the “diagnostic” role that measurement can have in such contexts. Rather than attempting to certify algorithms as af-




firmatively fair, we might instead develop measures to detect when they behave in normatively undesirable ways. Put simply, while we cannot hope to quantify fairness, we could at least detect egregious instances of unfairness. And to that end, so-called fairness metrics can be useful insofar as they can alert us to (and possibly mitigate) problematic behavior.


**Algorithmic fairness: A diagnostic baseline?** This suggests, then, that one goal of the algorithmic fairness community should be to produce a suite of diagnostic tools that detect undesirable behaviors in algorithms. Indeed, several such toolkits are publicly available from companies such as IBM, Microsoft, and Google. Of course, no such toolkit can be complete; each application and domain will have its own idiosyncrasies, requiring more nuance than simply running a suite of predefined tests. But perhaps a toolkit could provide at least a baseline, or a minimum standard to which algorithms can be held.

Unfortunately, recent results tell us that even this is too much to hope for: Even simple, intuitive measures of fairness cannot be simultaneously achieved.<sup>10,25</sup> If we use these measures to detect problematic behavior, our algorithms will always appear problematic according to at least one of them. Thus, any attempt to create a context-agnostic baseline for fair algorithmic decision-making will fail simply because any algorithm is doomed to violate our intuitive norms. Moreover, it is important to note that these results are mathematical, as opposed to computational, in nature. It is not a question of computational hardness, but rather one of mathematical impossibility. Consequently, these theorems apply to all decision-making processes, whether algorithmic or human.

What, then, can we do to promote fair decision making? This article will explore the consequences of these impossibility results. First, we present the impossibility theorems in the context of risk-assessment tools in the criminal justice system, where they were initially formalized. Next, we survey extensions of these results throughout the computer science literature, including further impossibility theorems and relaxations that admit algorithmic solutions. We then consider proposals to navigate the



## When algorithms are used to make or assist with consequential decisions, how do we ensure they do so fairly?



trade-offs present in these theorems, spanning a variety of disciplines, including computer science, philosophy, and legal studies. Then, we discuss arguments that impossibility theorems should point us toward changes in the broader contexts in which algorithms are deployed, including their objectives and how they are used.

### Impossibility Theorems for Fair Decision Making

In May 2016, the news outlet *ProPublica* published a widely circulated article claiming to have uncovered racial bias in an algorithm used in the criminal justice system.<sup>3</sup> The algorithm in question, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), was developed by a firm called Northpointe, Inc. (now Equivant). COMPAS was deployed in pre-trial hearings to help judges decide whether to release defendants on bail or incarcerate them until their trials. Based on a defendant's responses to a survey, COMPAS was designed to predict several outcomes, including the likelihood that, if released, the defendant would engage in new criminal activity and fail to appear for subsequent court dates. COMPAS assigned scores between 1 and 10 to defendants for each outcome, where 1 signified the lowest likelihood and 10 signified the highest.

Obtaining data from Broward County, FL, journalists attempted to determine whether the predictions matched future outcomes and whether they exhibited racial bias. The most salient and widely cited aspects of their analysis centered around predictions of recidivism, or whether a defendant would be arrested for new criminal activity. Angwin et al.<sup>3</sup> discovered worrying disparities in the rates at which COMPAS made different types of errors for white and Black defendants. Of those who ultimately were not re-arrested, Black defendants were given significantly higher scores than their white counterparts. We can think of these higher scores as akin to a higher false positive or Type 1 error rate for Black defendants. The same trend held for false negatives: Black defendants who went on to be re-arrested were still given significantly higher scores than analogous white defendants. *ProPublica*'s findings seemed damning.

According to the analysis, COMPAS assigned unjustifiably higher scores to Black defendants, which appeared to be clear evidence of racial bias. Facing mounting public scrutiny, Northpointe issued a formal response that claimed to refute *ProPublica's* claims.<sup>14</sup> Northpointe's analysis claimed that the measures that *ProPublica* had used (error rates for Black and white defendants) were the wrong ones; instead, Northpointe set out to show that COMPAS exhibited "accuracy equity" and "predictive parity."<sup>14</sup> Others in the actuarial risk community concurred, claiming to provide a "correct analysis" countering *ProPublica's* claims.<sup>17</sup> And while these two reports took issue with *ProPublica's* methodology, the basic contention still held: Whether or not they were arrested for subsequent crimes, Black defendants received higher scores than white defendants. The debate centered not on the facts or the data, but on the "right" way to measure racial bias in risk assessment.

As this debate continued, my co-authors and I sought to formalize these competing claims in a common mathematical framework in order to compare them.<sup>25</sup> In our framework, we consider a classification setting, where each individual belongs to either the positive class or the negative class (for example, recidivists and non-recidivists), and the goal of the risk assessment tool is to predict the likelihood that an individual is in the positive class. In the simplest version of the framework, each individual belongs to one of two groups. In the case of COMPAS, these groups would correspond to white and Black defendants. And finally, each individual comes with some feature vector  $\sigma$  providing information with which the risk assessment tool makes its predictions. We assume the risk assessment tool outputs the predicted probability  $p$  that each individual belongs to the positive class.<sup>a</sup>

In this framework, *ProPublica's* findings could be generalized as follows: Conditioned on belonging to the negative (or positive) class, COMPAS assigned higher scores to Black defendants than

to white defendants. Intuitively, this feels problematic; given that a defendant will (or will not) go on to recidivate, the algorithm's prediction of their risk should not depend on race. This suggests that normatively, we believe a risk-assessment algorithm should satisfy the following properties:

► **S.1: Balance for the negative class**—Conditioned on belonging to the negative class, white and Black defendants have similar scores.

► **S.2: Balance for the positive class**—Conditioned on belonging to the positive class, white and Black defendants have similar scores.

The problem, according to *ProPublica*, was that COMPAS did not respect these intuitive definitions of fairness. Northpointe, on the other hand, did not share this conception of fairness. Instead, it focused on whether COMPAS had similar predictive power for defendants of different racial groups. While its analyses considered a variety of statistical techniques, they all rested on a basic property: Conditioned on a prediction, the defendant's race provides no further information about the outcome of interest.<sup>14</sup> Often known as sufficiency or calibration, this property requires that predicted probabilities accurately reflect true outcomes:

► **S.3: Calibration within groups**—For each group and each predicted probability  $p$ , of the individuals who are in that group and assigned a prediction of  $p$ , a  $p$  fraction of them belong to the positive class.

In other words, calibration requires that predictions are statistically honest. Of the defendants assigned a predicted risk of, say, 15%, we can conclude that (roughly) 15% of these defendants belong to the positive class. And moreover, this holds regardless of race. Predictions of 15% can't mean 17% for white defendants and 13% for Black defendants; they must mean 15% for each racial group.

With these three statements, we can formalize the debate: *ProPublica* argued that COMPAS violated S.1 and S.2, while Northpointe's analysis relied on the fact that COMPAS satisfied S.3. At first glance, these statements may appear unremarkable, uncontroversial, and perhaps unrelated. But in fact, they allow us to get at heart of the COMPAS controversy through the following theo-

rem, which my co-authors and I proved in Kleinberg et al.<sup>25</sup>

**Theorem 2.1.** Unless one of the following conditions holds, statements S.1, S.2, and S.3 cannot simultaneously hold.<sup>25</sup>

► The individuals can be perfectly classified into the positive and negative classes (that is, the feature vectors  $\sigma$  convey perfect information).

► The fraction of individuals in the positive class is the same in both groups.

In other words, unless a predictor is perfect or two demographic groups have equal base rates (where the base rate is the fraction of individuals in the group who belong to the positive class), at least one of S.1, S.2, or S.3 must be violated. In concurrent work, Chouldechova<sup>10</sup> proved a similar result in a binary classification setting. This result also holds approximately: If statements S.1, S.2, and S.3 hold approximately, then the dataset must either admit approximately perfect prediction, or the fraction of individuals in the positive class must be approximately the same in the two groups. As one might expect, even approximately perfect classification is impossible in this setting; Northpointe's own analysis of the data released by *ProPublica* data reports an AUC of approximately 0.7.<sup>14</sup> And in that same data, the base rates of measured recidivism differ between white and Black defendants, a point highlighted by Dieterich et al.<sup>14</sup>

These results offer an explanation, though not necessarily a resolution, of the story. In light of Theorem 2.1, COMPAS could not have satisfied all of these conditions. Given that Northpointe made predictions that were relatively well calibrated within each racial group (consistent with the existing standards of the actuarial risk community), it could not simultaneously equalize error rates, leading to the disparities that *ProPublica* pointed out. On the other hand, some of the authors of the original *ProPublica* article countered that error-rate disparities could be reduced "if the algorithms focus on the fairness of outcomes," suggesting that Northpointe's choice to prioritize calibration was just that—a choice.<sup>2</sup>

And while it might be tempting to view Theorem 2.1 as an argument against the use of algorithmic risk prediction altogether in the criminal jus-

a While COMPAS simply produced a number between 1 and 10, these can be converted to probabilities by simply taking the fraction of defendants with each score who are in the positive class.

tice system, the impossibility is mathematical, not computational, in nature. According to Theorem 2.1, no decision maker, algorithmic or otherwise, can avoid violating at least one of these three conditions. The problem lies not with algorithmic risk prediction, but with the very act of predicting risk.

Risk prediction in the criminal justice system was central to the *ProPublica* story, providing a clear, intuitive example of how fairness is normatively desirable. But these impossibility theorems apply beyond the criminal justice context. For decision-support algorithms used in domains such as employment, housing, and financial services,<sup>5</sup> these same theorems apply. Attempting to predict qualification for employment, housing, or creditworthiness across populations with different distributions will lead to either miscalibration (violating S.3) or error-rate disparities (violating S.1 or S.2). While the remainder of this article will primarily discuss risk prediction in the criminal justice system, similar concerns arise in these other domains as well.

In the years that have followed these impossibility theorems, scholars and practitioners have built on them in several directions, which we examine here. We begin by considering some of the technical extensions of this work, including further impossibility theorems and technical formulations that seek to find a middle ground in these trade-offs. Next, we discuss the practical implications of these results. The difference in perspectives from *ProPublica* and Northpointe previews the wide range of responses that various researchers and communities offered to the inherent conflict between seemingly desirable normative goals. Then we will consider some of these responses and discuss their implications for algorithmic decision making in practice.

## Extending the Results

**Further impossibility.** Theorem 2.1 demonstrates the impossibility of simultaneously satisfying multiple constraints that appear normatively desirable. Of course, these are not the only criteria that we may want to enforce algorithmically. One line of work has seen scholars formulate other desirable properties in a variety of contexts and similarly show that these properties conflict with each

other. Corbett-Davies and Goel<sup>11</sup> show that efforts to achieve classification parity—a set of properties including S.1 and S.2—are hampered by legal and normative pressures toward anti-classification or directly relying on demographic attributes in decision making. Similarly, Narayanan<sup>31</sup> and Berk et al.<sup>6</sup> define several of measures of fairness based on the standard two-by-two confusion matrix and prove that various combinations of these measures conflict with one another. Even for mutually compatible measures, Pleiss et al.<sup>32</sup> show that achieving them in practice requires strictly degrading performance for one demographic group while failing to improve outcomes for any others, effectively requiring Pareto-suboptimality.

Other work has sought to investigate tensions between different notions of fairness beyond the binary decision-making context. One such context is representation learning, where one seeks to learn a representation  $Z$  for a dataset  $X$  such that  $Z$  contains as much predictive information as possible about outcomes of interest  $Y$  while minimizing information about a demographic attribute  $A$ . Thus, regardless of how a downstream decision maker uses  $Z$ , the lack of information about  $A$  will guarantee that any predictions made from  $Z$  will behave similarly across demographic groups. In this setting, Lechner et al.<sup>26</sup> show it is impossible to construct such a representation  $Z$  in a task-independent way; representations must inherently be sensitive to the task and data distribution at hand.

Another important context for fair decision making is resource allocation, where a decision maker must determine how to fairly distribute scarce resources. Using measures of fairness such as S.1 and S.2, multiple studies have pointed out both theoretical and empirical trade-offs between equity and utility in resource allocation.<sup>15,22</sup> Conceptually, these works highlight the tension between concentrating resources toward those most likely to be in need and ensuring that resources are available to all.

**Searching for relaxations.** While these further impossibility results have expanded our understanding of what cannot be done, other studies have sought to find relaxations that achieve some middle ground between competing measures. Of course, these relax-

ations cannot circumvent the underlying tensions, but they can help us better understand exactly what our normative goals are and how we might approximate them, even as they remain out of reach. For example, Reich and Vijaykumar<sup>36</sup> consider a two-step decision process (like Rambachan et al.<sup>35</sup>), where an algorithm provides a continuous output to which a human decision maker applies a threshold. They show that it is possible for the algorithm to be well calibrated (satisfying S.3) while enforcing that the overall “thresholded” decisions result in equal error rates across demographic groups (satisfying S.1 and S.2). While this requires that the human decision maker behave as simply a threshold function, which may not be admissible in practice,<sup>b</sup> it probes at the question of why, and at what stage in the decision-making process, we may want calibration in the first place. The two-stage decision model lends itself to further relaxations in which a classifier is given the option to defer decisions. Madras et al.<sup>29</sup> and Canetti et al.<sup>8</sup> consider models where the goal is to design a binary classifier that satisfies some definition of fairness across different groups, but the classifier may defer some decisions to a human decision maker. Formally, in addition to outputting 0 and 1 (corresponding to the negative and positive classes respectively), the classifier may also output  $\perp$ , meaning the decision will pass to a human. In these models, it is theoretically possible to design classifiers that satisfy any combination of desired criteria on the subset of data for which they do not output  $\perp$ , bypassing impossibility theorems. To understand why, recall from Theorem 2.1 that all three criteria can hold when the fraction of individuals in the positive class is the same in both groups. Suppose that the classifier could identify a subset of the instances such that both groups the same fraction of individuals in the positive class. Then, a classifier applied to this subset (that is, outputting  $\perp$  for all other individuals) could satisfy the three criteria S.1, S.2, and S.3. Of course, achieving this may be technically challenging, as studies show.<sup>8,29</sup> Moreover, while the classifiers themselves may bypass im-

<sup>b</sup> See, for example, *Loomis v. Wisconsin*, 881 N.W. 2d 749 (Wis. 2016).

possibility theorems by only acting on a subset of the data, the overall decision process (including the human decision maker to whom some decisions are deferred) must still be subject to the original impossibility theorems. Another relaxation approach is to view the trade-offs between fairness criteria as a multi-objective optimization problem. Several authors take this view, providing both theoretical and empirical results on managing trade-offs.<sup>9,19</sup> Efforts like these enable us to trace out Pareto-curves between different criteria, providing a richer picture of the tensions between them.

A further direction involves questioning the accuracy of the underlying data on which a model is evaluated. In this view, the data-generating process may be biased, meaning accuracy and fairness with respect to the observed data do not imply the same with respect to the ground truth. Blum and Stangl<sup>7</sup> explore this perspective, finding that under certain assumptions about biased data generation, constraints such as S.1 and S.2, which are intended to promote fairness, can in fact increase a decision-maker's accuracy. Thus, fairness need not conflict with accurate decision making; accounting for existing biases in the world can advance both fairness and accuracy.

### Navigating Trade-offs

Despite ongoing interest in designing algorithms to produce fairer decisions, no amount of technical innovation can circumvent the fundamental issue: We cannot design decision-making systems that satisfy all the properties we want them to. As long as inequities persist in society, decision-making systems (algorithmic or otherwise) will continue to reflect them. But a simple acknowledgement of inequity yields little information about how we should make decisions going forward. Should we accept the status quo and the error-rate disparities that come with it? Do we strive to implement changes, in our algorithms or more broadly, that seek to break the cycle of inequality? Perhaps predictably, scholars have articulated a wide range of views on these questions. As Narayanan<sup>31</sup> points out, this is in part because different measures of fairness reflect the perspectives of different stakeholders: A decision maker who is primarily



**Despite ongoing interest in designing algorithms to produce fairer decisions, no amount of technical innovation can circumvent the fundamental issue: We cannot design decision-making systems that satisfy all the properties we want them to.**



interested in accuracy may focus on calibration, while a decision subject may be concerned with being wrongly labeled high risk, or the false positive rate. Here, we survey and discuss responses from the academic literature and beyond.

**A focus on maintaining calibration.** The actuarial risk community has primarily settled on calibration as a requirement for risk assessment tools.<sup>14,17</sup> In this view, the goal of a tool is to provide accurate information to a human decision maker, not to try to correct for injustices that have led to observed disparities between demographic groups and the resulting error-rate differences.<sup>18</sup> Moreover, some practitioners and scholars contend that calibration is not just an end in itself, but also a necessary condition for other important measures of fairness. Drawing upon standards for educational testing laid out by the American Psychological Association, Skeem and Lowenkamp<sup>38</sup> emphasize the importance of properties such as accuracy equity (that is, whether a predictor's accuracy is similar when measured across different demographic groups), which are not well defined for poorly calibrated predictors. In fact, prioritizing calibration to ensure accuracy equity is a common theme across responses to the *ProPublica* article from the actuarial risk community.<sup>14,17,38</sup>


Another argument against using error-rate disparities to measure fairness comes from infra-marginality, as Corbett-Davies and Goel<sup>11</sup> argue. In this view, our understanding of fairness should depend on what happens at the margin, where small changes in the decision rule or policy would change outcomes. For example, if a decision maker applies a threshold to a calibrated continuous risk score (including, perhaps, the “true” risk distribution), then individuals whose scores fall near the threshold would be at the margin. Measures of error rates (such as S.1 and S.2), however, are sensitive to changes far from the margin. For example, dramatically increasing the number of high-risk individuals may change the false positive rate, even though the decision rule itself has not changed. Corbett-Davies and Goel<sup>11</sup> argue that this makes error-rate disparities a poor indicator of unfairness. Instead, they advocate for a utility-based framework that accounts for the costs and benefits of various ac-




tions, for which calibration is usually necessary (but not sufficient).

Advocating for calibration in algorithms is not necessarily in tension with interventions designed to promote equity. Some researchers advocate for the separation of algorithm development from actual decision making, arguing that algorithms should be made as accurate as possible, and decision-makers' preferences for equity should determine how algorithmic outputs are used.<sup>35</sup> While impossibility theorems would still apply to the overall decision process, according to this perspective, algorithm development is the wrong place to debate or express preferences for fairness. Instead, if a decision maker seeks to promote equity in decision making, they should explicitly do so through how they use algorithmic predictions, as opposed to through how they construct predictions.

**Grounding in philosophical principles.** Philosophers have also tried to provide normative reasons to prefer one conception of fairness or another. Loi and Heitz<sup>27</sup> develop a moral argument according to a particular definition of fairness, arguing that whether fairness requires calibration depends on whether one group is made worse off than another as a result. Long takes a stronger position in favor of calibration from the perspective of procedural fairness, arguing that while procedural fairness requires calibration, disparities in error rates are irrelevant from this perspective.<sup>28</sup> Similarly, Hedden<sup>23</sup> makes the case that measures of fairness based on error-rate disparities are morally irrelevant, and calibration is the only measure that has a bearing on fairness. He argues that measures of error-rate disparities, including S.1 and S.2, can be violated by an algorithm that makes optimal predictions given the available information and is not influenced by group membership, even when two populations have equal base rates.<sup>23</sup> By his definition, such an algorithm is not unfair, meaning that equalizing error rates cannot be a necessary condition for fairness. Importantly, each of these arguments begins with normative principles and derives conclusions from them. While these initial principles are still subject to debate, they provide concrete beliefs under which some measures of fairness are preferable to others, help-



**Beyond downstream impacts on individuals, constraining error-rate disparities creates positive incentives for decision makers.**



ing to formalize debates between different measures.

**Reducing outcome disparities.** While calibration seeks to ensure decisions accurately reflect past outcomes, it makes no attempt to account for the potential impacts that these decisions may have in the future. To the extent that data encodes past bias or discrimination, a calibrated predictor will perpetuate these biases in decisions going forward. In this view, calibration on its own does not lead to fair decision making; the evaluation of a predictive system must be sensitive to its downstream consequences.

Chouldechova<sup>10</sup> makes this concrete by modeling the consequences of a failure to enforce S.1 under the simple assumption that being labeled high risk is more costly to an individual than being labeled low risk. She demonstrates that disparities in error rates translate to higher costs borne by individuals in the disadvantaged group—in the case of COMPAS, Black defendants.<sup>10</sup> Srebro makes a similar point, saying disparities in error rates imply that the disadvantaged group is “paying the price for the uncertainty” of the algorithm.<sup>2</sup>

Beyond downstream impacts on individuals, constraining error-rate disparities creates positive incentives for decision makers. As Hardt et al.<sup>21</sup> write, preventing error-rate disparities “helps to incentivize the collection of better features, that depend more directly on the target rather than the protected attribute, and of data that allows better prediction for all protected classes.” Constraining error-rate disparities thus not only shifts costs away from disadvantaged groups in the short term, but also creates an environment in which predictions are less likely to exhibit racial disparities in the future.

Hellman<sup>24</sup> also makes the case for reducing error-rate disparities from both normative and legal perspectives. She argues that calibration is fundamentally concerned with beliefs, not actions. By definition, a calibrated prediction is one that accurately forecasts the probability of an event occurring. In other words, we can use a calibrated prediction to form an accurate belief about the world, but such a belief does not directly translate to an action. According to Hellman, this makes calibration “ill-suited as a measure of fairness,”<sup>24</sup> since fair-

ness should be concerned with actions, not beliefs. Instead, she advocates for a particular measure of outcome disparities—specifically, the ratio of false positives to false negatives—as a significant measure, since it conveys how a decision maker balances different types of costs. If this ratio differs across demographic groups, this suggests the decision-maker’s priorities are not the same for these groups, a potential indicator of unfair decision-making.

**Soliciting stakeholder views.** Given this apparent disagreement about which metrics matter and when, some scholars have proposed methods to solicit stakeholder opinions on what to prioritize when these metrics conflict with one another. For example, Srivastava et al.<sup>39</sup> design experiments to elicit non-expert opinions on what constitutes fair decision making. Yu et al.<sup>40</sup> develop tools to help algorithm designers explore the different possible trade-offs between measures of fairness, showing that they can help users better express their own values. Studies such as these seek to understand whether the measures of fairness in question are truly important to stakeholders, and if so, how they think about the ensuing trade-offs.

There is no consistent agreement that one of these measures must be prioritized over the others; moreover, context matters. A measure that seems relevant in one context may not express a meaningful societal value in another. And while debates over the meaning of different ways to measure disparities can be instructive, some scholars argue that a focus on measuring the properties of algorithms detracts from important broader conversations about the social systems in which these algorithms are deployed. We turn to this class of perspectives next.

### Reconsidering Broader Contexts

Given the generality of the impossibility theorems presented here, it may be tempting to view them as fundamentally unavoidable. Indeed, given that a decision maker is committed to making predictions about risk, any decision process, algorithmic or otherwise, will be subject to them. But why must we accept this framing, that decisions must be based on forecasted risk? If impossibility results are binding within certain predictive contexts, instead of

trying to tweak our algorithms to manage trade-offs within these contexts, we might instead change the contexts themselves to allow for decision making that is more just. By reconsidering the broader contexts in which decisions are made, we may be able to make decisions more fairly while sidestepping formal impossibility results. Here, we survey two directions in this spirit, with the understanding that depending on the specific domain and constraints in question, there is far more room for creative solutions than we can cover in this article. In particular, we consider proposals to re-frame how predictions are used and push beyond fairness for substantive reform.

**Rethinking prediction.** Machine learning is designed for predictive tasks. Given labeled data, we have increasingly sophisticated tools to predict labels on new instances. But the fact that we can predict does not necessarily mean that we should predict. Why should a judge’s decision to incarcerate a defendant be based on a prediction of their future behavior? Why should a hiring decision depend on a prediction of an applicant’s productivity? Not every problem must become a nail just because we are holding a hammer; some problems may not be amenable to prediction. And even when predictions may be useful, they need not be used to directly influence decisions.

The case for rethinking the role of prediction is not a new one. Scholars from a variety of disciplines have argued that in certain contexts, prediction should not be used at all. For example, Harcourt<sup>20</sup> advocates against prediction in policing and the criminal justice system, arguing that punishment should depend on what a person has done, not on what they are likely to do. In this view, predictions about the future are irrelevant; past behavior should be the only determinant of punitive actions, and only to the extent that it merits punishment. Even though predictions are derived from past behavior, they do not reflect normative judgements on the merits of that behavior, making them inadmissible from this perspective.

Richardson et al.<sup>37</sup> advance this view in an amicus brief to the Pennsylvania Supreme Court, in part drawing on impossibility theorems. They advocate against the use of actuarial risk assess-

ment in Philadelphia’s bail system, writing, “[n]o technical methodology can overcome the fact that different racial fairness metrics are currently irreconcilable.”<sup>37</sup> In this way, impossibility results can serve as a means for “rebuttal”: Establishing limits on what can be achieved through prediction “can expose the limitations of an entire category of approaches”<sup>1</sup> and provide pressure to adopt other frameworks for decision making.

And yet, actuarial assessments are commonplace in a wide range of contexts, including criminal justice. Recognizing this, scholars have begun to articulate ways to use these predictions in ways that make violations of fairness measures less salient. The goal of these efforts is not to somehow circumvent impossibility theorems, but to re-frame the types of decisions being made, casting the inevitable disparities that arise from prediction as an obligation to provide support and resources to those deemed “high risk.” In effect, these efforts focus less on the predictions themselves and more on what we choose to do with them.

In the criminal justice system, particularly when predicting arrest, data from the past is riddled with discriminatory practices, some of which have since been declared unconstitutional.<sup>30</sup> Thus, Mayson argues, the act of prediction introduces distortions in decision-making processes.<sup>30</sup> Impossibility theorems imply that these distortions cannot be fixed by algorithmic means alone, and as a result, Mayson<sup>30</sup> contends that we should rethink how we interpret predictions of risk. In particular, she argues against conflating risk with blame; we should view a judgement of high risk as an indicator that an individual needs additional support as well as a tool to target the structural conditions that lead to high risk.<sup>30</sup>

Thus, predictions need not be used to make decisions about reward and punishment; instead, they can serve to inform and target interventions toward those in need. Barabas et al.<sup>4</sup> develop this position from a causal inference perspective. Instead of simply predicting outcomes, they argue that causal inference tools can help us understand why certain individuals or groups appear to be high risk and, as a consequence, what sort of interventions



might help mitigate those risks without imposing punitive outcomes.<sup>4</sup> Of course, these data-driven predictions or inferences will still be subject to impossibility theorems like Theorem 2.1: Any predictions of need for support will either suffer from miscalibration or error-rate disparities when base rates differ. However, we may find violations of these conditions less normatively problematic when they apply to interventions designed to help, rather than punish, decision subjects.

**Substantive reform.** Expanding the scope of decision-support algorithms beyond prediction may help to dull, if not completely remove, the impact of impossibility theorems. However, some scholars argue that this does not go far enough. According to this view, real justice requires a broader reckoning with the environments in which decisions are made as well as the scope of changes we are willing to consider. Green<sup>18</sup> contends that the focus on individual decisions is already too narrow; “substantive algorithmic fairness,” as he terms it, requires us to acknowledge existing injustices and affirmatively seek to remedy them. Green’s perspective is that the focus on the impossibility of fair decision making is misguided, and substantive algorithmic fairness enables us to “escape from the impossibility of fairness.”<sup>18</sup>

To do so, Green<sup>18</sup> puts forward an approach that centers on substantive inequalities, as opposed to algorithms. In this approach, we must diagnose the relevant inequalities and identify what reforms might remedy these inequalities before we determine how (if at all) algorithms can play a positive role.<sup>18</sup> Measures of fairness are thus only useful insofar as they serve the broader goal of addressing systemic inequality. And if fairness measures are not inherently meaningful, impossibility theorems between them do not prevent us from pursuing substantive reform. Davis et al.<sup>13</sup> take a related view in advocating for what they term “algorithmic reparation,” a framework which seeks to explicitly recognize and redress structural inequities through algorithmic decision making. They contend that algorithmic systems that learn from structurally biased data will replicate these biases; beginning to remedy them requires an algorithm designer

to actively consider structural inequities and attempt to undo them in future decisions. Like Green, Davis et al.<sup>13</sup> view debates over calibration, error-rate disparities, and other fairness measures as beside the point. The framing of “fairness” fails to capture this broader commitment to highlighting injustices and working to remediate them.

Concretely, Davis et al.<sup>13</sup> give the example of an algorithm used to make hiring decisions in the tech sector, drawing inspiration from a widely circulated story about a hiring tool developed by Amazon that learned to penalize women applicants (through proxies like women’s clubs or colleges) from past data in which men were predominantly selected.<sup>12</sup> An approach based on the measures of fairness considered in this paper would focus on whether these predictions were well calibrated and on the error rates they produced. In contrast, the “reparative” approach would explicitly give higher scores to “women, trans, and non-binary applicants”<sup>13</sup> in recognition of the social conditions that have led these groups to be underrepresented in existing data to begin with. And while this is motivated in part by the desire to redress past injustice, it has a forward-looking component as well: A “reparative system would literally value the contributions underrepresented applicants bring to the company”<sup>13</sup> by recognizing their contribution to a more just workplace.

In a sense, we can interpret this as a call to reconsider the objective of algorithmic decision making. Algorithms are often designed to predict specific, measurable quantities (for example, whether an applicant would be hired) simply because those are the measures on which we have data. But in practice, a decision maker can have other objectives, such as hiring from the local community, fostering inclusivity, and redressing past discrimination. Standard measures of fairness lack the nuance to express these objectives, but attentiveness to underlying social structures can lead to algorithmic interventions that explicitly seek to further them.

Still, there are political and operational limits to implementing these frameworks in practice. As Green<sup>18</sup> points out, answering questions about social hierarchies and reforms that

might alter them “can be a difficult and politically contested task.” Disagreements about what constitutes a just decision-making process must be resolved somewhere, and the frameworks provided by Green<sup>18</sup> and Davis et al.<sup>13</sup> seek to make them explicit. When decisions are made by the government, as in the criminal justice context discussed throughout this article, this sort of political debate can occur through a democratic process. In the private sector, however, it is somewhat less clear how we should expect a for-profit company to assess and remedy structural inequities.

From an operational perspective, companies may not have the expertise to conduct the kind of in-depth social analysis needed to understand how their actions might reduce or exacerbate social hierarchies. This may require “new practices and training for computer scientists,”<sup>18</sup> or perhaps employees from more diverse disciplinary backgrounds. And while this may be desirable, a corporation may not be incentivized to make these investments compared to a public institution with a direct obligation to social welfare.

## Discussion

Fairness is contested. As a society, we do not agree on what it means. But we might have once hoped that we could at least operationalize the bare minimum that we do agree on. This turns out not to be true. There are intuitive, reasonable conditions on which we broadly agree but cannot achieve under realistic conditions.<sup>10,25</sup> In this article, we have summarized the theorems that formally establish these impossibility results and surveyed their extensions and implications.

One line of work has extended these results to prove further impossibility theorems for fair decision making in a variety of contexts, including classification, representation learning, and resource allocation. Others have constructed technical relaxations in which weaker or more restricted versions of the conditions discussed in this article can be achieved.

In parallel, impossibility theorems for fair decision making have sparked debate over which measures of fairness are truly desirable and what society should prioritize. No consensus has emerged from these debates; dif-

ferent philosophical and legal traditions reach different conclusions about what constitutes unfairness and how to measure it.

For some, the tension observed in impossibility results is an artifact of the framing of making fair predictions, and escaping from impossibility theorems requires that we change the broader contexts in which algorithms are deployed. We might eschew predictions of risk and instead focus on interventions designed to support those in need, re-framing the predictive question of “What will this person do?” as “Why is this person predicted to be high risk, and what support will reduce or eliminate that risk?” It can also mean explicitly seeking to redress structural inequities in decision making, such that an algorithm can be normatively evaluated not on some rigid measure of fairness but instead by the extent to which it supports the goal of reducing inequity.

From these lines of work, a few themes have emerged. First, there is an interpretation of these results reminiscent of “no-free-lunch” theorems in machine learning. In an unequal world, algorithmic tools cannot simply make that inequality disappear. A decision maker who wants to enforce some idea of fairness must take an affirmative stand on what that entails. They cannot hope to remain agnostic by constraining a system to be fair by all possible measures, since it is impossible to do so.

Second, it has become clear that context matters significantly. Our normative ideas about fairness differ, from criminal justice to medical decision making to employment, and these differences shape how we think about fairness trade-offs. A one-size-fits-all approach to fairness cannot succeed. Attempts to make decision making fairer must attend to the social, legal, and practical constraints of a given setting. Finally, debates over fairness raise questions about the scope of interventions that we are willing to consider. A focus on algorithms and their properties may detract from efforts toward broader reform. Instead of assuming the contexts in which data-driven decision making is deployed are fixed, we might instead use formal impossibility results to push for more fundamental change in the types of decisions that

are made. This raises deep normative and political questions about what sort of change is feasible and who should make those decisions.

Ultimately, normative judgements about what constitutes fair decision making cannot be automated or delegated to algorithms. Impossibility theorems, such as the ones surveyed in this article, help us to formalize this idea: Disagreements about fairness do not simply disappear when algorithms are involved. And while measures of fairness can provide some insight as to how decisions are made, they cannot on their own tell us how decisions should be made. **□**

#### References

- Abebe, R. et al. Roles for computing in social change. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2020), 252–260.
- Angwin, J. and Larson, J. Bias in criminal risk scores is mathematically inevitable, researchers say. *Ethics of Data and Analytics*, Auerbach Publications (2016), 265–267.
- Angwin, J., Larson, J., Mattu, S. and Kirchner, L. Machine bias. *Ethics of Data and Analytics*, Auerbach Publications (2016), 254–264.
- Barabas, C. et al. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2018), 62–76.
- Barocas, S., Hardt, M. and Narayanan, A. *Fairness and Machine Learning* (2019); <http://www.fairmlbook.org>.
- Berk, R. et al. Fairness in criminal justice risk assessments: The state of the art. In *Proceedings of Sociological Methods & Research* 50, 1 (Feb. 2021), 3–44.
- Blum, A. and Stangl, K. Recovering from biased data: Can fairness constraints improve accuracy? In *Proceedings of the 1st Symp. on Foundations of Responsible Computing 156* (2020), 3:1–3:20.
- Canetti, R. et al. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2019), 309–318.
- Celis, L.E., Huang, L., Keswani, V. and Vishnoi, N.K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2019), 319–328.
- Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *Proceedings of Big Data 5*, 2 (2017), 153–163.
- Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv* (2018); <http://arxiv.org/abs/1808.00023>.
- Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, Auerbach Publications (2018), 296–299.
- Davis, J.L., Williams, A. and Yang, M.W. Algorithmic repair. In *Proceedings of Big Data & Society* 8, 2 (2021).
- Dieterich, W., Mendoza, C. and Brennan, T. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc.* 7, 4 (2016).
- Donahue, K. and Kleinberg, J. Fairness and utilization in allocating resources with uncertain demand. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2020), 658–668.
- Dwork, C. et al. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf.* (2012), 214–226.
- Flores, A.W., Bechtel, K. and Lowenkamp, C.T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks. *Federal Probation J.* 80, (2016), 38.
- Green, B. Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology* 35, 90 (2022)
- Gultchin, L. et al. *Beyond Impossibility: Balancing Sufficiency, Separation and Accuracy* (2022); <http://arxiv.org/abs/2205.12327>.
- Harcourt, B.E. *Against Prediction*, University of Chicago Press (2008).
- Hardt, M., Price, E. and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of Advances in Neural Information Processing Systems* 29, (2016).
- He, Y., Burghardt, K., Guo, S. and Lerman, K. Inherent trade-offs in the fair allocation of treatments. *arXiv* (2020); <http://arxiv.org/abs/2010.16409>.
- Hedden, B. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs* 49, 2 (2021).
- Hellman, D. Measuring algorithmic fairness. *Virginia Law Rev.* 106, 4 (2020), 811–866.
- Kleinberg, J., Mullainathan, S. and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conf.* (2017).
- Lechner, T., Ben-David, S., Agarwal, S. and Ananthakrishnan, N. Impossibility results for fair representations. *arXiv* (2021); <http://arxiv.org/abs/2107.03483>.
- Loi, M. and Heitz, C. Is calibration a fairness requirement? An argument from the point of view of moral philosophy and decision theory. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* (2022), 2026–2034.
- Long, R. Fairness in machine learning: Against false positive rate equality as a measure of fairness. *J. of Moral Philosophy* 19, 1 (Nov. 2021), 49–78.
- Madras, D., Pitassi, T. and Zemel, R. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Proceedings of Advances in Neural Information Processing Systems* (2018).
- Mayson, S.G. Bias in, bias out. *Yale Law J.* 128, 8 (June 2019).
- Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency* 1170, 3 (2022).
- Pleiss, G. et al. On fairness and calibration. In *Proceedings of Advances in Neural Information Processing Systems* (2017).
- Powles, J. and Nissenbaum, H. The seductive diversion of “solving” bias in artificial intelligence. *OneZero, Medium* (2018).
- Raghavan, M. The societal impacts of algorithmic decision-making. *Ph.D. Dissertation*, Cornell University (2021).
- Rambachan, A., Kleinberg, J., Mullainathan, S. and Ludwig, J. An economic approach to regulating algorithms. *Technical Report*, National Bureau of Economic Research (2020).
- Reich, C.L. and Vijaykumar, S. A possibility in algorithmic fairness: Can calibration and equal error rates be reconciled? In *Foundations of Responsible Computing* (2021).
- Richardson, R., Schultz, J.M. and Jacobson, S.E. Brief for amicus curiae AI Now Institute in support of petitioners. *Philadelphia Community Bail Fund, et al. Pets v. Commonwealth Arraignment Ct.* 21 EM 2019 (Pa. 2020).
- Skeem, J. and Lowenkamp, C. Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law* 38, 3 (2020), 259–278.
- Srivastava, M., Heidari, H. and Krause, A. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining* (2019), 2459–2468.
- Yu, B. et al. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conf.* 1245–1257.

**Manish Raghavan** (mragh@mit.edu) is the Drew Houston Career Development Professor at the MIT Sloan School of Management and Department of Electrical Engineering and Computer Science, Cambridge, MA, USA.

© 2024 copyright held by the owner/author(s).  
Publication rights licensed to ACM.