Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms

Lily Morse[a], Mike Horia M. Teodorescu[b,c], Yazeed Awwad[c], & Gerald Kane[b]

[a]West Virginia University, College of Business & Economics

lily.morse@mail.wvu.edu

[b]Boston College, Carroll School of Management

teodores@bc.edu

kanegb@bc.edu

[c]Massachusetts Institute of Technology, D-Lab

awwad@mit.edu

This version of the article has been accepted for publication at the *Journal of Business Ethics* (doi: 10.1007/s10551-021-04939-5). The present document does not reflect post-acceptance improvements, or any corrections.

Abstract

Recent advances in machine learning methods have created opportunities to eliminate unfairness from algorithmic decision making. Multiple computational techniques (i.e., algorithmic fairness criteria) have arisen out of this work. Yet, urgent questions remain about the perceived fairness of these criteria and in which situations organizations should use them. In this paper, we seek to gain insight into these questions by exploring fairness perceptions of five algorithmic criteria. We focus on two key dimensions of fairness evaluations: distributive fairness and procedural fairness. We shed light on variation in the potential for different algorithmic criteria to facilitate distributive fairness. Subsequently, we discuss procedural fairness and provide a framework for understanding how algorithmic criteria relate to essential aspects of this construct, which helps to identify when a specific criterion is suitable. From a practical standpoint, we encourage organizations to recognize that managing fairness in machine learning systems is complex, and that adopting a blind or one-size-fits-all mentality toward algorithmic criteria will surely damage people's attitudes and trust in automated technology. Instead, firms should carefully consider the subtle yet significant differences between these technical solutions.

KEYWORDS: fairness; machine learning; distributive fairness; procedural fairness; algorithm design

Do the Ends Justify the Means?: Exploring Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms

Machine learning (ML) is appealing for organizations because it reduces tedious tasks and can enhance decision performance in situations where human bias and errors are likely (Miller, 2018; Pezzo & Beckstead, 2020; Silverman & Waller, 2015). As such, managers increasingly rely on ML algorithms to make decisions. Yet the rise of artificial intelligence tools has also sparked new ethical challenges for business and society (Greenwood et al., 2020; Kim & Scheller-Wolf, 2019; Leicht-Deobald et al., 2019; Martin, 2019a; North-Samardzic, 2019).

One crucial challenge involves ensuring that ML models make decisions that are fair and inclusive. This area of research is active in computer science. It has led to the development of many statistical techniques, known collectively as fairness criteria, that embed notions of fairness into the design of algorithms. However, researchers have primarily conducted this work without considering people's perceptions of these criteria. Consequently, we lack an understanding of whether individuals believe they are fair—an important predictor of people's willingness to trust and support algorithmic decisions as well as the organizations who implement them (McFarlin & Sweeney, 1992; Newman et al., 2020). It is further unclear whether relevant differences exist in how people perceive these metrics. If found, such variation can be used to inform when managers and developers should apply a particular criterion, if at all, as they are often mutually incompatible.

To address these questions, we explore the perceived fairness of five algorithmic criteria proposed in the computer science literature. Using an organizational justice theory lens, we focus on two key elements that shape individuals' fairness perceptions—distributive fairness (i.e., the fairness of decision outcomes) and procedural fairness (i.e., the fairness of decision processes).

Our analysis leads to several insights. First, we shed light on variation in the potential for different algorithmic criteria to facilitate distributive fairness. More broadly, we discern that statistical solutions for fairness tend to emphasize distributive fairness concerns. Subsequently, we discuss procedural fairness and propose a framework for understanding how algorithmic criteria relate to essential aspects of this construct, which helps to identify when a specific criterion is suitable. From a practical standpoint, our research might motivate changes to the way managers and developers oversee ML systems. Rather than adopting a blind or one-size-fits-all mentality toward existing fairness criteria, which will surely damage people's attitudes and willingness to accept algorithmic decisions, we advise practitioners to carefully consider the subtle yet significant differences between these technical solutions.

Background of Fairness in ML

ML is the field of computer science referring to *algorithms* – a set of machine-computable instructions that solve a problem in a finite number of steps – which derive patterns from prior data. ML is at the intersection of computer science, statistics, linguistics, and mathematics as a research field. The key objective of ML is to enable predictions that improve as additional data becomes available to the algorithm. While ML tools are promising in providing substantial increases in organizational efficiency and cheaper implementation of specific tasks, significant shortcomings should not be overlooked, including those that may negatively impact fairness. Only recently did fairness and ethics issues begin to take a more prominent role in ML scholarship by emphasizing that developers should expect an approach of "awareness" to the problem of fairness as part of the development process (Dwork et al. 2012). This view has given rise to the subfield of fairness in ML. However, it is essential first to understand how ML models operate in practice.

ML involves the use of computer algorithms to create models from data automatically. These algorithms learn from existing labeled datasets where developers label the observed outcomes for every predictor variable, otherwise known as a *feature* (James et al., 2013). For example, a bank may have data profiles of applicants who filed for a mortgage application and the bank's decision for each applicant. The developer then splits this labeled dataset into a *training set* and a *test set*. This process allows the ML model to train or tune its parameters to fit the data but leaves out part of the known (labeled) data to verify whether the algorithm produces correct predictions (Teodorescu, 2017). For a binary criterion variable, such as hiring a job candidate or not, the developer can categorize the outcome as follows: true positive (the organization hires the candidate), true negative (the organization rejects the candidate), false positive (the algorithm predicts the organization will hire the candidate, but it rejects the candidate), and false negative (the algorithm predicts the organization will reject the candidate, but it hires the candidate). The developer then uses these four values to determine the accuracy of the algorithm.

Accuracy represents the ratio of correct predictions (actual hires, actual rejections) versus the total number of prediction attempts. Although accuracy is the primary and often default measure used by programmers to determine the performance of an algorithm, it does not capture any of the nuances in type I or type II errors (for an overview of types of classification errors and costs of such errors in different fields, see Martin, 2019b). More broadly, accuracy is conceptualized solely in terms of predicted outcomes based on a test set – comparing the correct responses to predicted responses and counting the correctly predicted against total attempts.[1]

It is noteworthy that designing and managing algorithms from a perspective of accuracy says virtually nothing about notions of *fairness*, defined in computer science as lack of "any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" (Mehrabi et al., 2019, p. 1). Algorithms that maximize the model's prediction accuracy may behave differently toward different subgroups within the data, leading to misclassification errors and unfair bias. For instance, a teacher may be unfairly deemed a poor worker by an ML model due to age, leading to their wrongful dismissal (O'Neil, 2016). A more well-known example is the recidivism prediction system COMPAS which discriminated against defendants based upon their race and gender (Brennan et al., 2009).

In response to these prediction pitfalls, the subfield of fairness in ML has focused on engineering algorithms that mathematically incorporate fairness ideals while maintaining a level of accurate performance. Many fairness criteria have arisen from this work (Hardt et al., 2016; Teodorescu & Yao, 2021). Each criterion provides a narrow definition of fairness, as it must for the sake of formalism, which are not all satisfiable concurrently. In other words, there is no one universally accepted conceptualization of fairness in ML (Verma & Rubin, 2018). The present article discusses five of the most popular fairness criteria in computer science: fairness through unawareness, demographic parity, accuracy parity, equality of opportunity, and equalized odds.

[1] Actual human operators label the outcomes in the case of supervised learning. This paper assumes that the developers train the ML model on a training set created using human input.

Unlike more inscrutable ML, researchers operationalize these definitions in terms of their tradeoffs between fairness and performance accuracy (Martin, 2019b).

The most straightforward criterion is fairness through unawareness, where variables deemed sensitive to unfairness, such as gender, age, ethnicity, and disability status, are dropped from the prediction model. In theory, this approach would make the ML model unaware and unable to discriminate based upon sensitive characteristics. However, as some of these variables tend to highly correlate with other features in the data that do end up in the model or deduced from other variables (the issue of "redundant encoding"; Bird et al., 2019), this approach simply does not work well. It may end up perpetuating discrimination while its overseers are unaware of it (Hardt et al., 2016).

Moving beyond this criterion, which is the same as not checking for discrimination by the algorithm, there are 'fairness-aware' approaches, four of which we focus on here: demographic parity, accuracy parity, equalized odds, and equality of opportunity. While each of these criteria presents unique tradeoffs in fairness versus accuracy, they all broadly seek to ensure fairer outcomes across protected and non-protected groups.

A Note on Protected Attributes

It is important to note that in addition to utilizing fairness criteria, a separate critical step in developing fairer automated systems involves determining which features in a training dataset constitute *protected attributes.* Protected attributes represent demographic features such as race, gender, age, sexual orientation, disability status, marital status, ethnicity, national origin, and socioeconomic status. If a feature in the dataset represents a protected attribute, developers should never use it as a predictor in the ML model.

The computer science literature has historically relied upon legally protected characteristics when determining what qualifies as a protected attribute. In the United States, these protected characteristics are codified into law through equal opportunity in hiring (FEEO), credit lending (ECOA), non-discrimination based on gender or race (Civil Rights Act Title VII 1964), and non-discrimination based on disability (ADA 1990, Rehabilitation Act 1973). Though legal safeguards and laws may only capture a small strand of characteristics that people believe merit protection in a given situation, this is a sensible starting point as the fairness field in computer science is relatively new. Indeed, the legal field had implemented the concept of a protected class long before ML existed, representing different categories within a given protected attribute. The idea has expanded over time through acts of Congress in the United States (as were the laws mentioned above) and through legal scholarship debate (Clarke, 2017; Schwartz, 2000). As such, it was perhaps the most convenient method to use an already defined set of protected categories from the legal literature to begin testing for fairness in a newer field such as ML.

Questions remain regarding whether this is an ethical or socially desirable approach to fairness. For example, laws usually take longer to negotiate or litigate than creating new technology; hence there is often a lag between what companies may do in practice or what is considered fair by society versus what is deemed lawful. Who should decide whether someone is part of a protected class is important and has real consequences for individuals' livelihood. Currently, the court system mainly settles this issue (Clarke, 2017). In the field of ML, however, there is no universally accepted solution for how developers should determine whether certain features or ambiguous cases in the data should be protected beyond existing protected attributes. Furthermore, the current lawmaking system and litigation for selecting protected attributes do not have input from ethicists or computer scientists on what should qualify as a protected class. We discern that this is an opportunity to change the status quo and that business ethics is integral.

Algorithmic Fairness Criteria: Insights from Organizational Justice Theory

Business ethicists sit at the crossroads between business, technology, and society (Martin & Freeman, 2004). We integrate theory from organizational justice scholarship to provide a deeper understanding of algorithmic fairness criteria with this consideration in mind. This knowledge is critical for theoretical reasons as well as practical ones. The central theoretical problem is that organizations and developers must choose between the algorithmic criteria, as they are mutually incompatible. Yet, we do not know how to people will react to the use of a particular criterion. The central practical problem is that the advice we can offer to organizations who wish to implement algorithmic criteria is lacking, potentially creating friction with employees, customers, partners, and broader communities (Lee, 2018; Newman et al., 2020).

Organizational Justice Theory

Organizational justice theory is broadly concerned with people's perceptions of fairness in the workplace, including distributive and procedural components (c.f. Colquitt, 2012; Colquitt & Rodell, 2015; Goldman & Cropanzano, 2015; Greenberg, 2011; Khan et al. 2015).[2, 3]

---

[2] Organizational justice researchers have also studied fairness as a single dimension (e.g., Ambrose & Schinke, 2009) and as a multidimensional construct comprising distributive, procedural, and interactional fairness (Colquitt et al., 2013; Karriker & Williams, 2009).

[3] In line with organizational justice scholarship, we use the terms fairness and justice interchangeably. Although differences exist among the concepts, both are geared toward promoting equity and avoiding bias.

Distributive fairness refers to the perceived fairness of outcomes and is judged according to how fair a decision is in its effect on the distribution of rewards and resources (Adams, 1965; Colquitt et al., 2001). Procedural fairness reflects the perceived fairness of decision processes, including how developers make decisions (e.g., Are procedures consistent? Are decisions based on accurate and bias-free information?) It also reflects how much control individuals have over the decision process (e.g., Are there opportunities for correcting flawed decisions?) (Leventhal, 1980; Thibaut & Walker, 1975).

In the sections that follow, we describe the distributive and procedural fairness of five popular computational solutions for resolving bias in ML. In doing so, we identify a central theme across this research: algorithmic criteria tend to emphasize distributive fairness concerns. Informed by insights from organizational justice theory, we subsequently explore whether procedural fairness can be enhanced and consider the role of contextual influences in shaping justice experiences. We specify some situations in which developers might increase perceptions of procedural justice for each fairness criterion, which provides a foundation for understanding when a given fairness metric may be suitable.

Algorithmic Criteria Emphasize Distributive Fairness

Although our understanding of organizational justice research in the ML landscape is nascent, computer science scholars are beginning to recognize that they have designed fairness metrics to focus almost exclusively on distributive fairness (Saxena et al., 2019; Selbst et al., 2019). Indeed, a recent review by Robert et al. (2020) noted that researchers operationalize technical definitions of algorithmic fairness regarding the equity of the outcomes received, which involves comparing one's inputs to obtained outputs relative to others. Figure 1 describes five fairness criteria that are among the most popular in computer science and the extent to which they achieve distributive fairness ideals. Together these metrics represent two broad approaches to fairness that have received substantial attention in the literature: a blindness approach (i.e., fairness through unawareness) and a group-focused approach (i.e., ensuring equality across one or several measures for all categories of a protected attribute).

Within the group-focused category, we further discern that researchers design some metrics to achieve parity ideals (i.e., an equal distribution of outcomes among subgroups despite differences, such as demographic parity and accuracy parity). At the same time, they develop other metrics to achieve equity ideals (i.e., an equal distribution of opportunities based on the circumstances of each subgroup, such as equality of opportunity and equalized odds). For simplicity, we focus on two metrics for each subtype in this paper. We expect that our critical analysis generalizes to other criteria that fall within these subtypes as they are considered the same from a distributive fairness perspective.

Importantly, we observe that some computational solutions, such as fairness through unawareness, likely achieve relatively low levels of distributive fairness. In contrast, others adopt a more proactive approach and thus achieve increasingly fairer outcomes. Yet, as indicated by the y-axis in the figure, each deeper intervention requires more significant technical effort and comes with a greater risk of over-correcting and forcing equality where it is not expected (Dwork et al. 2012).

Blindness Approach: Fairness Through Unawareness

The most commonly applied approach in organizations is fairness through unawareness. Developers consider this metric "unaware" such that it will simply ignore fairness information by leaving out protected attributes from the data such as age, sex, and race/ethnicity. It is not surprising that developers and managers have favored this technique seeing that organizations have historically adopted similar approaches for managing diversity and equality. For instance, the colorblind diversity strategy, defined by a belief that organizations should treat people equally no matter their cultural background, is still used in many occupational settings and involves denying or not "seeing" race or other sensitive attributes (Apfelbaum et al. 2010; Podsiadlowski et al. 2013). Although colorblind ideologies may appear to function successfully on the surface, thereby promoting an illusion of fairness in the short-term, research indicates they are ineffective in rooting out perceived bias and instead perpetuate social inequities over time (Ely & Thomas 2001). Indeed, ignoring the plausibility of discrimination often results in stronger perceptions of unfairness and worse outcomes for members of minority groups (Purdie-Vaughns & Eibach, 2008).

Fairness through unawareness likewise fails to reduce discrimination and prejudicial outcomes in practice. A critical flaw of this criterion is that ignoring differences due to different backgrounds does not change the fact that other variables in an ML model may strongly correlate with protected attributes or sensitive features removed from the data. These correlations effectively serve as proxies for the removed variables, making a mockery of the claim to be "unaware".

For example, the algorithm used to determine credit lines for an Apple credit card did not include gender as an input yet learned to rely on inputs highly correlated with gender, such as historical salary data that contained hidden prejudices against women. Public responses to the algorithm's credit lending decisions were numerous and hostile, as evidenced by Twitter profiles of affected applicants. Even Apple's co-founder Steve Wosniak raised fairness concerns, questioning "whether the card might harbor some misogynistic tendencies" (Knight, 2019, para. 9). Taken together, we contend that fairness through unawareness does little to ensure that individuals will perceive a fair distribution of rewards and resources. It, therefore, achieves an unacceptably low level of distributive fairness.

Group-Focused Parity Approach: Demographic Parity and Accuracy Parity

Demographic parity is a well-known fairness intervention in which the ML algorithm reaches a positive outcome at the same rate irrespective of the categories of a protected attribute. For example, if a firm's hiring rate for one gender is 20%, then the hiring rate for all other values of gender should also be 20% irrespective of other constraints. This approach is more accountable than fairness through unawareness because the developer makes a conscious decision to tune the algorithm. In the hiring scenario, the developer ensures that the positive outcome (e.g., a recommended hire) is independent of gender. Thus, the sensitive variable is not discarded but rather a part of the process of ensuring equitable outcomes (Kusner et al., 2017). From an organization justice standpoint, demographic parity promotes greater distributive fairness than does fairness through unawareness. Particular challenges remain, however, that restricts this approach's promise in real-world settings.

Demographic parity is concerned with preventing adverse or disparate impacts for disadvantaged groups, yet a significant downside is that it often fails to reach fair outcomes in practice. In particular, demographic parity cannot deal with differences between subgroups other than to assume that success rates are equal. In other words, it neglects individual unfairness, sacrificing in some cases qualified individuals to obtain equality at the group level. Consider a hiring scenario in which an organization wishes to achieve equal hiring success rates across two groups, group A and group B. Candidates from group A tend to be less qualified than the least-qualified candidate in group B. If demographic parity is applied, the organization would hire from the two groups at the same rate; however, this would likely create perceptions of unfairness for the qualified but rejected candidates in group B. In this example, demographic parity might initially lead management to form impressions that hiring outcomes are fairer due to the parity created across the two groups. Yet, these judgments may soon fade as people begin to realize that the model overlooks qualified applicants from one particular group.

This outcome can also engender negative attitudes toward candidates who would not otherwise have qualified but were hired to meet the parity requirement, potentially leading to a self-fulfilling prophecy. Imagine a company that rigorously hires male job applicants at a rate of 35% and indiscriminately hires female applicants at the same rate (Ghassami, 2018). Although the acceptance rate in both gender groups is the same, the low effort to ensure that the algorithm chooses the best female candidates under demographic parity will likely cause female hires look like poor performers. The result may establish a negative track record for the female group.

Like any simple solution to bias in ML, demographic parity is by design a blunt instrument and tends to fix one distributive fairness problem at the cost of exacerbating others. Notably, demographic parity can compound outcome-based bias against those who are members of multiple sensitive categories. For example, while developers can enforce equality across genders, as can equality across races, there is no protection against the possibility that such methods will exacerbate bias against specific gender-race pairings or other combinations of sensitive attributes (Teodorescu et al., 2021). Such errors in fairness intensify as the data becomes more imbalanced across sensitive groups, including multiple-protected groups. Thus, we propose that demographic parity may obtain, at best, perceptions of moderately fairer outcomes (i.e., a moderate level of distributive fairness) in practice depending on the data and parameters under which the algorithm must optimize.

A closely related cousin to demographic parity is accuracy parity. As previously discussed, the default algorithm performance measure in ML is accuracy, operationalized as the ratio between the count of correctly predicted outcomes to the overall count of prediction attempts. While this measure does not distinguish between Type I and Type II errors, optimizing an algorithm is intuitive and straightforward. An extension of this measure to fairness would involve subsetting the data by the protected attribute and calculating the accuracy per subgroup (Zhao et al., 2020). In this case, the algorithm is considered fair in computer science if the subgroups' accuracy is equal (or close). Accuracy parity is well-liked by computer scientists because accuracy is often the default measure in standard ML packages. Researchers calculate accuracy by running the trained model onto the test set to infer the expected behavior of the model out of the sample.

However, like demographic parity, several constraints are associated with accuracy parity that inhibit the ability to achieve fairer outcomes, producing moderate perceptions of distributive fairness in occupational settings. First, we may be trading off the false positives of one group for another group's false negatives and not know about it. Second, accuracy parity works poorly for datasets where the classes are imbalanced (i.e., there is no even division across subgroups). As an extreme example, let us assume we have a dataset where the model rejects 95% of job applicants, irrespective of other attributes. A classifier that simply returns "reject" for all applicants would have an accuracy of 95% and pass accuracy parity. This outcome, of course, would not be perceived as fair to the applicant pool. Thus, we need more sophisticated approaches that rely on more than just accuracy.

Group-Focused Equity Approach: Equality of Opportunity and Equalized Odds

Equality of opportunity measures whether individuals who should qualify for an opportunity have the same likelihood of being deemed qualified by the ML model regardless of the value of a protected attribute. Like demographic and accuracy parity, this fairness metric focuses on equalizing positive outcomes. Specifically, it ensures that, whatever the value of the protected attribute, the model equalizes the rate of a predicted positive result for a qualified individual (Hardt et al., 2016; Kusner et al., 2017). Thus equality of opportunity is a more targeted metric that allows for demographic differences but levels the playing field by requiring that unfair/erroneous judgments, or false-positive rates, be equitably distributed. In an organizational hiring example, while strong male and female candidates may be of comparable quality, the algorithm could detect a significant difference among the weak candidates across gender. In this case, *weak* corresponds to job performance for the firm's position, such as females in the weak job performance category being more qualified candidates than males.

The equalized odds fairness metric is a stricter version of equality of opportunity. It adds to the requirements that the true positive rate and the false positive rate are equal across categories of the same protected attribute (Hardt et al., 2016). If false positive rates are significantly different between two categories, the one with lower false positives may feel that the outcomes are biased and thus feel demeaned. Essentially, instead of having just one equality condition, such as true positive rate equal across sample subgroups by protected attribute, we must now satisfy a system of two equations. It involves strict equality (in the pure mathematical definition) across both true positive and false positive rates across all population subgroups by protected attributes. The requirement of a system of equations and strict equality makes this criterion more challenging to fulfill.

Because equality of opportunity and equalized odds take a more nuanced and narrow approach to ensuring equitable outcomes across different groups, we suggest they potentially produce higher perceived levels of distributive fairness. By the same token, they tend to address very targeted forms of outcome-based unfairness, which results in a scenario where plugging one leak could result in the emergence or worsening of other leaks.

Overall, reflecting on these five criteria, we broadly conclude that computer scientists have made rapid progress in engineering algorithms that incorporate distributive fairness concerns. However, there is still no known roadmap for applying these metrics so that people will perceive algorithmic decisions as fairer. This leads to the question of where to go from here. How can managers and organizations determine whether a particular criterion is the "right" one? To answer this question, we draw attention to the role of procedural fairness, which we argue can provide a more robust understanding of when individuals will perceive a specific fairness criterion as suitable.

Procedural Fairness and Algorithmic Fairness Criteria

Procedural fairness, defined as the perceived fairness of the methods used to make decisions (Colquitt, 2001), plays a vital role in shaping how people react to decisions. Six components underlie procedural fairness: consistency, accuracy, ethicality, representativeness, bias suppression, and correctability (Leventhal, 1980). Consistency reflects in the uniformity of decision procedures across people and time. Accuracy represents the extent to which methods utilize valid, high-quality information. Ethicality captures whether practices uphold moral standards and values. Representativeness demands that procedures duly consider the needs and concerns of the entire group. Bias suppression requires that decision procedures are impartial and prevent favoritism by the decision-maker. Lastly, correctability captures techniques that provide opportunities to challenge or correct flawed decisions.

Procedural fairness is essential because "just processes signal that [individuals] are valued and esteemed by their referent social groups" (Cropanzano & Stein, 2009, p. 200). Procedures also signal the decision maker's goals, such as intentions to maximize societal welfare, which provides vital information about why the decision-maker made a particular choice (Tyler, 2003). While distributive and procedural fairness mutually influence justice evaluations, procedural fairness has been considered the more robust predictor of the two (Thibaut & Walker, 1975; van den Bos et al., 2001). Indeed, people are more willing to support an unfair outcome when they feel the process is fair. They especially rely on procedural fairness when information about a decision-maker’s trustworthiness is uncertain (van den Bos et al., 1998), which is often the case for machine systems (Glikson & Woolley, 2020).

Conversations about procedural fairness in ML are beginning to emerge across the fields of computer science and management; however, this work has almost entirely focused on comparing the procedural fairness of algorithms to humans (e.g., Bigman et al., 2020, Lee, 2018; Newman et al., 2020). To date, we still know very little about the procedural fairness of algorithmic fairness criteria and whether relevant differences exist in how people perceive them (for exceptions, see Grgić-Hlača et al., 2018 and Lee et al., 2019). Yet if managers or developers only consider distributive fairness, as is essentially the case in computer science, they risk applying the wrong criterion at the wrong time and damaging people's attitudes and trust toward automated technology as well as the firms that implement it. Thus we encourage scholars and practitioners to develop a deeper understanding of procedural fairness in this area.

Procedural Fairness and the Role of Contextuality

While there are multiple ways to enhance procedural fairness in ML, we call attention to the situational context's power to shape justice perceptions. We chose to focus on situations because managers do not always know an algorithm's design and data structure. Still, they may more readily take responsibility for learning the contexts in which algorithmic fairness criteria are likely to be viewed as procedurally fair.

We reason that if organizations or developers apply a fairness metric in the right setting, people will perceive it as procedurally fairer. Our discussion primarily draws from prior organizational justice research indicating that people react more positively to decisions when the situational context heightens the salience of procedural fairness components (Farrar et al., 2020; Mathur & Sarin Jain, 2020). For example, fairness evaluations are higher when contextual conditions signal a decision maker's normatively ethical goals to others (as opposed to productivity goals, which often violate societal moral standards). This condition satisfies the ethicality component of procedural fairness (Barrett-Howard & Tyler, 1986). Presumably, situations that signal multiple components will more effectively increase perceptions of procedural fairness.

Our assessment of the capacity for different algorithmic criteria to signal procedural fairness components is depicted in Table 1. These insights translate to specific contextual applications for each metric. For simplicity, we focus on diversity and inclusion scenarios to illustrate the value of our analysis. We intend to demonstrate proof of concept, not develop a comprehensive solution. Further, the situational examples we discuss may change according to prevailing notions of fairness in a particular society at a given time. Lastly, it is also important to note that our distinctions among these criteria derive from their general approach to achieving fairness (blindness, group-focused with emphasis on parity vs. equity). We expect that other metrics that fall within these categories are viewed similarly from a procedural fairness perspective.

Blindness Approach: Fairness through Unawareness

We begin with fairness through unawareness, which enforces willful blindness to protected attributes linked to unfairness. Currently, this method is the default approach for organizations and developers when implementing ML models. However, the prevalence of fairness through unawareness stems more from its technical practicality and capacity to promote the illusion of fairness by making decision processes remarkably *consistent* compared to human decision-making, yet it severely neglects the remaining procedural fairness components. Because this metric does little to satisfy multiple procedural fairness concerns, it is doubtful whether individuals would view it favorably in any diversity and inclusion context.

Consider the case of Amazon's now-disbanded recruitment model that reviewed job candidates' resumes (Dastin, 2018). Originally intended to filter through hundreds of resumes to select qualified candidates (using consistent procedures), Amazon's model took an unexpected turn when developers discovered it to have developed a bias against women. The bias emerged from the data used to train the ML algorithm, which consisted of actual resumes submitted to Amazon over ten years. Because most job applicants in the data pool were male, the program determined that men were more qualified than women. Specifically, the model learned to downgrade candidates who belonged to all-female extracurricular groups or had graduated from all-women's colleges—predictors highly correlated with gender.

In the Amazon incident, the use of fairness through unawareness made it challenging to determine which characteristics the ML model was optimizing on, including whether protected attributes were imputed using proxy information. It called into question whether it achieved *accuracy*, *representativeness*, and *bias suppression*. Likewise, the design of fairness through unawareness did not provide interpretable signals to the public about the ML model's *ethicality*, such as moral motives to hire more diverse talent. A likely result is that people struggled to assess whether the algorithm's decisions were procedurally fair. Fairness through unawareness further detracted from procedural fairness by failing to embed mechanisms for rectifying flawed choices (i.e., *correctability*). By maintaining willful blindness, there is no practical way to counter indirect discrimination and bias when it arises. Together we believe these shortcomings led to negative perceptions of fairness, which intensified as unfair outcomes inevitably emerged.

Accordingly, we argue that fairness through unawareness is rarely, if ever, suitable for practical use in diversity and inclusion contexts. Even if there is strong evidence that the algorithm's features do not correlate with the protected attributes—which is seldom, if ever, true—this approach cannot make procedural fairness components salient beyond consistency. Thus, people will likely perceive it as procedurally unfair. As long as algorithms are deployed in real-world settings and can influence individuals' fairness experiences, we caution against the use of fairness through unawareness.

That said, we acknowledge a possible exception to its preclusion. Namely, fairness through unawareness may be appropriate when organizations use ML models for purely mechanical, simple tasks. These tasks would not involve protected attributes in the data or directly impact human beings (e.g., using ‘hard-coded,’ rule-based systems). Examples include character recognition, object classification, and spam classification. Research has shown that people regard fairness through unawareness favorably in such situations (Lee, 2018), likely because the model's decisions to not personally affect individuals (otherwise known as low outcome dependence in the field of social psychology; van der Toorn et al., 2011).

Group-Focused Parity Approach: Demographic Parity and Accuracy Parity

Next, we explore the procedural fairness of demographic parity and accuracy parity. Given the technical constraints associated with these metrics in achieving fairer outcomes, it is important that organizations implement them in the right setting so that the procedural gains can outweigh potential distributive losses.

Like fairness through unawareness, demographic parity and accuracy parity can provide high *consistency* in decision procedures across persons and over time. Indeed, algorithms almost always surpass consistency levels achieved by human decision-makers (Lee et al. 2019). In contrast to fairness through unawareness, however, demographic parity and accuracy parity ensure that decisions are independent of protected attributes, yielding higher *accuracy*. These criteria are also designed with moral principles in mind, taking active steps to prevent disparate treatment and impact for disadvantaged groups. Thus, we argue they have higher potential to *suppress bias* as it arises in the decision process and deliver more interpretable signals to others about an ML model's *ethicality*. Ethicality may be especially easy to make salient in practice as managers and developers might only need to provide a basic understanding of these metrics to users to convey their moral intentions. On the other hand, the technical limitations of these criteria may make it tricky for practitioners to demonstrate that bias suppression has been achieved, and so we expect less strong effects on people's fairness judgments. Not to be overlooked, these criteria do little to fix poor or flawed decisions when they arise. Thus, they are likely interpreted as providing low *correctability*.

Perhaps most significantly, we contend that demographic parity and accuracy parity are unique amongst the metrics in their powerful potential to address concerns related to *representativeness*.[4] As part of their underlying structure, demographic parity and accuracy parity explicitly represent all affected subgroups and ensure equal rates of success between them. Accordingly, we suggest that organizations focus on applying these two criteria in situations where concerns for representativeness are serious. In doing so, the situational context is likely to heighten procedural fairness evaluations toward algorithms.

For example, managers who wish to hire more diverse talent may implement demographic parity as part of the automated stages of the job interview process. Managers should also explain their use of this criterion to job candidates. Because the structure of demographic parity can signal that members of underrepresented groups are valued and accepted, it may strengthen minority candidates' sense of belongingness and increase their trust in the ML model (Valcke et al., 2020). In some cases, there is intriguing evidence that deploying demographic parity or accuracy parity can attract more minority candidates to an organization and boost that minority group's social standing. For instance, Hu and Chen (2018) showed that applying demographic parity to hiring decisions combatted racial inequality in labor markets by

[4] Closely related to representativeness is the concept of voice, which allows individuals from different subgroups to express their concerns, opinions, and values to decision-makers as part of the decision process (Thibaut & Walker, 1975). While we believe that voice is an influential factor to consider when examining perceived fairness, it falls outside Leventhal's (1980) theory of procedural fairness criteria—our current focus here—and is thus beyond the scope of this paper.

temporarily increasing the number of minorities hired for entry-level positions. This result subsequently improved the societal reputation of that minority group and contributed to their downstream career success. Demographic parity and accuracy parity, then, may strongly enhance perceived fairness when properly applied.

Group-Focused Equity Approach: Equality of Opportunity and equalized odds

Finally, we turn to equality of opportunity and equalized odds, which we have argued potentially obtain fairer outcomes than the previous metrics. Given their distributive strengths, firms might initially conclude that these metrics are the front-runners for improving perceived fairness. However, they are also more challenging than others to implement in real-world settings due to their strict constraints. With these tradeoffs in mind, we examine the procedural fairness of these two criteria.

Similar to the other metrics discussed, we observe that equality of opportunity and equalized odds deliver highly *consistent* decision-making procedures due to their automated nature. They also do not take *correctability* into account in the event of bad outcomes or errors, at least not without human intervention. We further observe that equality of opportunity and equalized odds overlap with the parity-focused metrics in their capacity to signal *accuracy* and *ethicality*. First, these metrics ensure that decisions are not based on predicted attributes, which likely improves accuracy perceptions of the ML model. Second, they are designed to uphold standards of ethics and morality. For instance, equality of opportunity promotes equity ideals by ensuring that qualified individuals from different subgroups receive positive opportunities at the same rate. Equalized odds establishes similar equity between subgroups across both positive and negative outcomes.

Importantly, we argue that equality of opportunity and equalized odds outshine demographic parity and accuracy parity in their capacity to signal *bias suppression*. For one, these metrics apply more targeted rules to promote fair decision making (e.g., requiring that false-positive rates be equitably distributed) and thus better prevent preferential treatment from arising in the decision process. In the case of equalized odds, impressions of neutrality—a key aspect of bias suppression—may also increase because the decision rules influence everyone (Solomon et al., 2021). That is, all cases in the sample are affected across both positive and negative outcomes. Yet, these procedural advances carry some costs with respect to *representativeness*, as different subgroups can be treated unequally when equity principles are emphasized; therefore subgroups may not be represented in the same way.

Based on this analysis, we suggest that managers and developers might benefit most by deploying equality of opportunity and equalized odds in situations where bias and special treatment are serious concerns. We believe that procedural fairness evaluations may be particularly enhanced in such contexts. In diversity and inclusion scenarios, this may translate to removing barriers to entry. As an illustration, consider the American National Football League, which created the Rooney Rule in 2002 to reduce racial discrimination when hiring head coaches. The Rooney Rule requires football teams to interview at least two external minority candidates for head coaching positions and at least one minority candidate for other positions, including senior football operations, general managers, coordinators, and club presidents (Patra, 2020). Three years after the league implemented this policy, the percentage of Black coaches increased from 6% to 22% (Cook, 2021). We argue that applying equality of opportunity may similarly help to level the perceived playing field among job candidates in automated phases of the interview process. Organizations should explain their use of this criterion to both recruiters and job candidates, emphasizing its ability to reduce favoritism and close the societal gap between different groups.

We expect that equalized odds is more appropriate when people care about preventing bias for both positive and negative outcomes. Generally speaking, this concern arises less frequently in real-world settings. People tend to focus more on whether a decision obtained positive results across subgroups and pay less attention to whether adverse opportunities or outcomes are equal. This criterion also has stricter technical constraints that lead us to question its utility in practice. Yet, there are some institutions that may stand to benefit by applying equalized odds, at least at certain times and in certain situations. For instance, universities and colleges may want to implement equalized odds as part of the automated stages of the college admission process. Doing so may signal to the public that applicants from different backgrounds are equally likely to be considered if they are qualified (or denied if they are unqualified), thereby improving perceived fairness.

Discussion

This research has examined essential questions regarding the perceived fairness of five algorithmic criteria in ML, which has seen little integration with management and ethics research to date. Our main objective was to provide a more comprehensive understanding of how people may view the different criteria through the lens of distributive and procedural fairness, which provides a navigation aid for determining when a particular metric may be suitable. We shed light on variation in the ability for different algorithmic metrics to facilitate distributive fairness, noting that obtaining fairer outcomes comes at the cost of more significant technical effort. We also examine differences in the extent to which these criteria satisfy conditions of procedural fairness, which informs their contextual applications. In the spirit of interdisciplinary scholarship, we sought to provide a robust discussion of fairness that offers sufficient breadth and depth. Our analysis considers the complex interplay between human and machine, technology and organizations, processes and outcomes, and inherent tensions among fairness and accuracy across the different disciplines.

Theoretical Implications

Several theoretical implications arise from our discussion, which suggest future directions. First, the present work illuminates the potential for behavioral ethics research to enrich our theoretical understanding of ML tools. Looking forward, we encourage organizational behavior and ethics scholars to further explore the relevance of distributive and procedural fairness for ML algorithmic criteria as we still have much to learn. One direction for future work is to examine the extent to which our theorizing generalizes to categories of algorithmic criteria not covered in our conceptual analysis. For instance, the computer science literature has paid growing attention to metrics that emphasize individual fairness and causal reasoning, such as fairness through awareness, counterfactual fairness, and fairness in relational domains (Lazo, 2020). Thoughtful consideration of the ways in which these criteria connect to distributive and procedural fairness is needed.

In particular, we expect that certain metrics might have unique relationships with procedural fairness that extend beyond our discussion in this paper. Fairness in relational domains, for example, may have pronounced effects on perceived *accuracy* and *bias suppression.* This criterion takes a socially rich set of information (individual, relational, organizational, and ecological data) into account when making decisions (Farnadi et al., 2018). For instance, fairness in relational domains can holistically evaluate the managerial opinions of an employee in contexts such as performance reviews and promotions. Managerial opinions are collectively inferred from previous performance reviews while the metric simultaneously recognizes and prevents relational bias from by a particular manager from emerging.

As we did not perform empirical studies in this paper, we also encourage future scholars to test the arguments we have put forward in occupational settings. Further, scholars might productively build upon on our work by exploring other situational contexts that may improve procedural fairness perceptions of algorithmic criteria, beyond diversity and inclusion. For example, retailers may apply certain metrics in trade and marketplace settings to more fairly determine living wages of the people who produce the products for sale, which in turn may foster positive perceptions toward the ML model and the retailer.

Practical Implications

In addition to theoretical implications for scholars, we also offer recommendations for organizations that deploy ML systems. As evidenced by corporate mission statements and company ethics codes, many, if not most, organizations state that they value integrity and ethical conduct. Managers and developers may also personally care about improving fairness, such as those with high levels of moral character (Cohen et al., 2014; Cohen & Morse, 2014). For such actors, an obvious implication is to avoid adopting a blindness or a one-size-fits-all approach toward algorithmic fairness criteria. Instead, we advise practitioners to carefully consider the conceptual differences among these choices and select a metric that best aligns with the situation at hand when developing an ML model.

We also strongly recommend that practitioners widen their understanding of the variables that may be sensitive to unfairness within a particular dataset. While the current standard in computer science focuses on legally protected attributes, there are many more factors that contribute to people's worldviews of fairness. For example, management research has linked organizational tenure (Hambrick et al., 1996), functional background and values (Jehn et al., 1999), politics (Chao & Moon, 2005), physical appearance (Rafaeli & Pratt, 1993), attitudes and personality (Harrison et al., 1998), network ties (Beckman & Haunschild, 2002), and pay (Pfeffer & Langton, 1988) to fairness evaluations, though these are not legally protected. We appreciate that these characteristics are fluid and likely fluctuate across situations, time, and societies. Still, it is incumbent on organizations to make concerted and frequent efforts to discern which features should be protected when applying a fairness criterion.

Lastly, it is essential to note that while it would be tempting to simply deploy a fairness metric in a particular situation and, after the organization achieves some performance data, fail to oversee or maintain it, doing so would violate the *correctability* element of procedural fairness. Because correctability is noticeably absent throughout the algorithmic criteria we have assessed, organizations should offset this weakness by tasking humans with monitoring the ML model's decisions and stepping in when errors and bad outcomes arise (Teodorescu et al., 2021). Indeed, humans and machines must work together to alleviate unfairness in this new digital age of work. Mastering a suite of algorithmic fairness criteria and building cross-functional talent in management teams with ML and business ethics backgrounds would do much to resolve the challenges exemplified in our paper.

References

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 267-299). Academic Press.

Ambrose, M. L., & Schminke, M. (2009). The role of overall justice judgments in organizational justice research: A test of mediation. *Journal of Applied Psychology*, *94*(2), 491-500.

Apfelbaum, E. P., Pauker, K., Sommers, S. R., & Ambady, N. (2010). In blind pursuit of racial equality? *Psychological Science*, *21*(11), 1587-1592.

Barrett-Howard, E., & Tyler, T. R. (1986). Procedural justice as a criterion in allocation decisions. *Journal of Personality and Social Psychology*, *50*(2), 296-304.

Beckman, C. M., & Haunschild, P. R. (2002). Network learning: The effects of partners' heterogeneity of experience on corporate acquisitions. *Administrative Science Quarterly*, *47*(1), 92-124.

Bigman, Y., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). Algorithmic Discrimination Causes Less Moral Outrage than Human Discrimination. *PsyArXiv*. https://doi.org/10.31234/osf.io/m3nrp

Bird, S., Kenthapadi, K., Kiciman, E., & Mitchell, M. (2019). Fairness-aware machine learning: Practical challenges and lessons learned. *Proceedings of the ACM International Conference on Web Search and Data Mining*. https://dl.acm.org/doi/fullHtml/10.1145/3308560.3320086

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System. *Criminal Justice and Behavior*, *36*(1), 21-40.

Chao, G. T., & Moon, H. (2005). The cultural mosaic: A metatheory for understanding the complexity of culture. *Journal of Applied Psychology*, 90(6), 1128-1140.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv*. https://arxiv.org/abs/1810.08810

Clarke, J. A. (2017). Protected Class Gatekeeping. *NYU Law Review, 92*, 101.

Cohen, T. R., & Morse, L. (2014). Moral character: What it is and what it does. *Research in Organizational Behavior, 34,* 43-61.

Cohen, T. R., Panter, A. T., Turan, N., Morse, L., & Kim, Y. (2014). Moral character in the workplace. *Journal of Personality and Social Psychology, 107*(5), 943-963.

Cook, I. (2021). *How HR Can Tackle Diversity Using the Rooney Rule*. Visier. Retrieved from https://www.visier.com/clarity/how-hr-can-tackle-diversity-using-the-rooney-rule/

Colquitt, J. A. (2012). Organizational justice. In S. W. J. Kozlowski (Ed.), *The Oxford Handbook of Organizational Psychology* (pp. 526-547). Oxford University Press.

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425-445.

Colquitt, J. A., & Rodell, J. B. (2015). Measuring justice and fairness. In R. S. Cropanzano & M. L. Ambrose (Eds.), *Oxford Library of Psychology. The Oxford Handbook of Justice in the Workplace* (p. 187-202). Oxford University Press.

Colquitt, J. A., Scott, B. A., Rodell, J. B., Long, D. M., Zapata, C. P., Conlon, D. E., & Wesson, M. J. (2013). Justice at the millennium, a decade later: A meta-analytic test of social exchange and affect-based perspectives. *Journal of Applied Psychology*, *98*(2), 199-236.

Cropanzano, R., & Stein, J. H. (2009). Organizational justice and behavioral ethics: Promises and prospects. *Business Ethics Quarterly*, 193-233.

Dastin, J. (2018, October 10). *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*. Reuters Business News. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the Innovations in Theoretical Computer Science Conference*. https://arxiv.org/abs/1104.3913

Ely, R. J., & Thomas, D. A. (2001). Cultural diversity at work: The effects of diversity perspectives on work group processes and outcomes. *Administrative Science Quarterly*, *46*(2), 229-273.

Farnadi, G., Babaki, B., & Getoor, L. (2018). Fairness in relational domains. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. https://dl.acm.org/doi/abs/10.1145/3278721.3278733

Farrar, J., Massey, D.W., Osecki, E., & Thorne, L. (2020). Tax Fairness: Conceptual Foundations and Empirical Measurement. *Journal of Business Ethics, 162*, 487-503.

Ghassami, A. 2018. Fairness in Supervised Learning: An Information Theoretic Approach. *IEEE International Symposium on Information Theory*. https://doig.org/10.1109/isit.2018.8437807.

Glikson, E., & Woolley, A. W. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals, 14*(2), 627-660.

Goldman, B., & Cropanzano, R. (2015). "Justice" and "fairness" are not the same thing. *Journal of Organizational Behavior*, *36*(2), 313-318.

Greenberg, J. (2011). Organizational justice: The dynamics of fairness in the workplace. In S. Zedeck (Ed.), *APA Handbook of Industrial and Organizational Psychology* (pp. 271-327). American Psychological Association.

Greenwood, B. N., Adjerid, I., Angst, C., & Meikle, N. (2020). How Unbecoming of You: Online Experiments Uncovering Gender Biases in Perceptions of Ridesharing Performance. *Journal of Business Ethics*, *18*, 1-20.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. *Proceedings of the AAAI Conference on Artificial Intelligence*. http://mlg.eng.cam.ac.uk/adrian/AAAI18-BeyondDistributiveFairness.pdf

Hambrick, D. C., Cho, T. S., & Chen, M-T. (1996). The Influence of Top Management Team Heterogeneity on Firms' Competitive Moves. *Administrative Science Quarterly*, *41*(4), 659-684.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1610.02413

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface-and deep-level diversity on work group cohesion. *Academy of Management Journal*, *41*(1), 96-107.

Hu, L., & Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. *Proceedings of the World Wide Web Conference*. https://arxiv.org/abs/1712.00064

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly*, *44*(4), 741-763.

Karriker, J. H., & Williams, M. L. (2009). Organizational justice and organizational citizenship behavior: A mediated multifoci model. *Journal of Management*, *35*(1), 112-135.

Khan, K., Abbas, M., Gul, A., & Raja, U. (2015). Organizational justice and job outcomes: Moderating role of Islamic work ethic. *Journal of Business Ethics*, *126*(2), 235-246.

Kim, T. W., & Scheller-Wolf, A. (2019). Technological unemployment, meaning in life, purpose of business, and the future of stakeholders. *Journal of Business Ethics*, *160*(2), 319-337.

Knight, W. (2019, November 19). *The Apple Credit Card Didn't 'See' Gender—and That's the Problem.* Wired. Retrieved from https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1703.06856

Lazo, C. (2020). Toward Engineering AI Software for Fairness [MSc Thesis, Delft University of Technology]. *TUDelft Respository*.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1).

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction.* https://doi.org/10.1145/3359284

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. *Journal of Business Ethics*, *160*, 377-392.

Leventhal G. S. (1980) What Should Be Done with Equity Theory? In Gergen K. J., Greenberg, M. S., & Willis R. H. (Eds.), *Social Exchange* (pp. 27-55). Springer.

Martin, K. (2019a). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, *160*, 835-850.

Martin, K. (2019b). Designing Ethical Algorithms. *MIS Quarterly Executive, 18*(2), 129-142.

Martin, K., & Freeman, R. E. (2004). The separation of technology and ethics in business ethics. *Journal of Business Ethics, 53*(4), 353-364.

Mathur, P., & Sarin Jain, S. (2020). Not All That Glitters is Golden: The Impact of Procedural Fairness Perceptions on Firm Evaluations and Customer Satisfaction with Favorable Outcomes. *Journal of Business Research*, *117*, 357-367.

McFarlin, D. B., & Sweeney, P. D. (1992). Distributive and procedural justice as predictors of satisfaction with personal and organizational outcomes. *Academy of Management Journal, 35*(3), 626-637.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv*. https://arxiv.org/abs/1908.09635

Miller, A. P. (2018, July 26). Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review*. Retrieved from https://hbr.org/2018/07/want-less-biased-decisions-usealgorithms

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes, 160*, 149-167.

North-Samardzic, A. (2019). Biometric technology and ethics: beyond security applications. *Journal of Business Ethics*, *167*, 433-450.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Patra, K. (2020, May). *NFL Instituting Changes to the Rooney Rule*. NFL. Retrieved from https://www.nfl.com/ news/nfl-instituting-changes-to-rooney-rule

Pezzo, M. V., & Beckstead, J. W. (2020). Algorithm aversion is too often presented as though it were non-compensatory: A reply to Longoni et al. (2020). *Judgment and Decision Making*, *15*(3), 449-451.

Pfeffer, J., & Langton, N. (1988). Wage inequality and the organization of work: The case of academic departments. *Administrative Science Quarterly*, *33*(4), 588-606.

Podsiadlowski, A., Gröschke, D., Kogler, M., Springer, C., & Van Der Zee, K. (2013). Managing a culturally diverse workforce: Diversity perspectives in organizations. *International Journal of Intercultural Relations*, *37*(2), 159-175.

Purdie-Vaughns, V., & Eibach, R. P. (2008). Intersectional invisibility: The distinctive advantages and disadvantages of multiple subordinate-group identities. *Sex Roles*, *59*(5-6), 377-391.

Rafaeli, A., & Pratt, M. G. (1993). Tailored meanings: On the meaning and impact of organizational dress. *Academy of Management Review*, *18*(1), 32-55.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction*, *35*(5-6), 1-31.

Saxena, N., Huang, K., DeFilippis, E., Radanovic, G., Parkes, G., & Liu, Y. (2019). How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. *arXiv*. https://dl.acm.org/doi/pdf/10.1145/3306618.3314248?casa_token=GkHedlYHaFoAAAAA:09D1cVv-RSI6iDIyUMG0XG7KRY4CSOCZXHDaphcv0nfmBRlgEUrGKVAymSJs-OUMhcRP1lyDDJS8

Schwartz, D. S. (2000). The Case of the Vanishing Protected Class: Reflections on Reverse Discrimination, Affirmative Action, and Racial Balancing. *Wisconsin Law Review*, 657.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287598

Silverman, R. E., & Waller, N. (2015, March). *The algorithm that tells the boss who might quit*. Wall Street Journal. Retrieved from https://www.wsj.com/articles/the-algorithm-that-tells-the-boss-who-might-quit-1426287935

Solomon, B., Hall, M. E. K., & Muir, C. P. (2021). When and Why Bias Suppression is Difficult to Sustain: The Asymmetric Effect of Intermittent Accountability. *Academy of Management Journal.* Advance online publication. https://journals.aom.org/doi/abs/10.5465/amj.2020.0441

Teodorescu, M. (2017). Machine Learning methods for strategy research. *Harvard Business School Research Paper Series* #18-011. https://www.hbs.edu/faculty/Pages/item.aspx?num=53076

Teodorescu, M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human–ML Augmentation, *Management Information Systems Quarterly*. Advance online publication. https://10.25300/MISQ/2021/16535

Teodorescu, M., & Yao, X. (2021). Machine Learning Fairness is Computationally Difficult and Algorithmically Unsatisfactorily Unsolved. *Proceedings of IEEE High Performance Computing Conference*.

Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Erlbaum.

Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. *Crime and Justice*, *30*, 283-357.

Valcke, B, Van Hiel, A, Onraet, E, & Dierckx, K. (2020). Procedural fairness enacted by societal actors increases social trust and social acceptance among ethnic minority members through the promotion of sense of societal belonging. *Journal of Applied Social Psychology*, *50*, 573-587.

van den Bos, K., Lind, E. A., & Wilke, H. A. M. (2001). The psychology of procedural and distributive justice viewed from the perspective of fairness heuristic theory. In R. Cropanzano (Ed.), *Series in applied psychology. Justice in the workplace: From theory to practice* (p. 49–66). Lawrence Erlbaum Associates Publishers.

van den Bos, K., Wilke, H. A. M., & Lind, E. A. (1998). When do we need procedural fairness? The role of trust in authority. *Journal of Personality and Social Psychology, 75*(6), 1449-1458.

van der Toorn, J., Tyler, T. R., & Jost, J. T. (2011). More than fair: Outcome dependence, system justification, and the perceived legitimacy of authority figures. *Journal of Experimental Social Psychology, 47*(1), 127-138.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*. https://doi.org/10.1145/3194770.3194776

Zhao, H., Coston, A., Adel, T., & Gordon, G. J. (2019). Conditional learning of fair representations. *arXiv*. https://arxiv.org/abs/1910.07162

Figure 1. Distributive Fairness of Algorithmic Criteria and their Technical Effort.

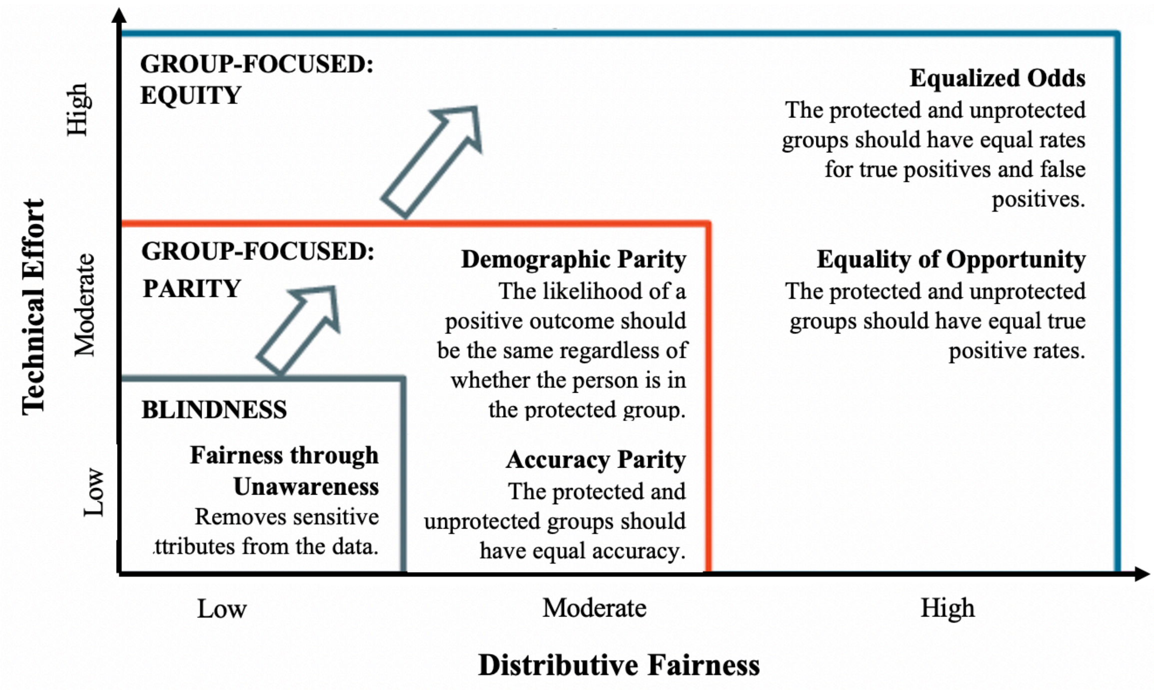

Figure 2. Framework of Algorithmic Criteria and their Relation to Procedural Fairness Components.

Fairness Metrics	**Ability to Signal Procedural Fairness Components**						**Contextual Applications: Diversity & Inclusion**
	Consistency	***Accuracy***	***Ethicality***	***Representativeness***	***Bias Suppression***	***Correctability***	
Fairness through Unawareness	High	Low	Low	Low	Low	Low	Rarely suitable.
Demographic Parity Accuracy Parity	High	Moderate	High	High	Moderate	Low	May be applied to improve representation of minority groups, such as hiring decisions.
Equality of Opportunity Equalized Odds	High	Moderate	High	Moderate	High	Low	May be applied to remove barriers to entry, such as in interview and college admissions processes.