What's Fair about Individual Fairness?

Will Fleisher

Pre-print version, forthcoming in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*,
DOI = 0 . 1145/ 3461702 . 3462621

Abstract

One of the main lines of research in algorithmic fairness involves individual fairness (IF) methods. Individual fairness is motivated by an intuitive principle, similar treatment, which requires that similar individuals be treated similarly. IF offers a precise account of this principle using distance metrics to evaluate the similarity of individuals. Proponents of individual fairness have argued that it gives the correct definition of algorithmic fairness, and that it should therefore be preferred to other methods for determining fairness. I argue that individual fairness cannot serve as a definition of fairness. Moreover, IF methods should not be given priority over other fairness methods, nor used in isolation from them. To support these conclusions, I describe four in-principle problems for individual fairness as a definition and as a method for ensuring fairness: (1) counterexamples show that similar treatment (and therefore IF) are insufficient to guarantee fairness; (2) IF methods for learning similarity metrics are at risk of encoding human implicit bias; (3) IF requires prior moral judgments, limiting its usefulness as a guide for fairness and undermining its claim to define fairness; and (4) the incommensurability of relevant moral values makes similarity metrics impossible for many tasks. In light of these limitations, I suggest that individual fairness cannot be a definition of fairness, and instead should be seen as one tool among several for ameliorating algorithmic bias.

1 Introduction

Algorithmic bias and fairness are becoming increasingly urgent topics as the use of AI and machine learning systems continues to spread through our society. As a result, research on algorithmic fairness has been expanding rapidly in the past few years.

One paradigm in algorithmic fairness research is *individual fairness* (IF). Proponents of IF suggest that the intuitive notion of fairness is expressed by the principle *similar treatment*: similar individuals should be treated similarly Dwork, Hardt, Pitassi, Reingold, and Zemel (2012). They attempt to capture this intuitive notion using two distance metrics: the first measures how similar

any two individuals are, and the second measures how similarly those individuals are treated. Individual fairness requires that if two individuals are close to each other according to the similarity metric, they be close to each other on the treatment metric.

Proponents of individual fairness argue that their theory "captures" the intuitive notion of fairness (Dwork et al. 2012, p. 214). They have suggested that similar treatment is the "guiding informal understanding of fairness" (Friedler, Scheidegger, & Venkatasubramanian 2016, p. 2). Moreover, they suggest that IF forbids types of unfairness that group fairness notions miss. For these reasons, they suggest individual fairness offers *the* formal definition of fairness. It should thereby have pride of place among methods for determining fairness and detecting bias.

I will argue for two primary claims. First, that individual fairness is inadequate as a definition of fairness. This is because similar treatment is insufficient to ensure fairness. Moreover, IF techniques require antecedent judgments about fairness, making IF unsuitable to provide a definition of the concept. Second, I will argue that IF should not be used as a sole means for determining whether an algorithm is fair, or for detecting bias. Doing so will be inadequate for ensuring fairness in a variety of circumstances. The upshot of my arguments is that IF is not specially justified as uniquely capturing the essence of fairness. Instead, IF should be considered one among several tools for algorithmic fairness.

I offer four arguments to support these conclusions:

1. **Insufficiency of similar treatment:** There are a wide variety of counterexamples to the sufficiency of this principle for ensuring fairness. This undermines the idea that IF provides a definition of fairness, or uniquely captures its essence. The insufficiency of similar treatment also suggests IF methods should not be used in isolation from other fairness methods.

2. **Systematic bias and arbiters**: One of the most promising ways of determining a similarity metric for individual fairness is to appeal to human arbiters to evaluate similarity of individuals. Arbiters provide feedback on whether and to what degree individuals are similar, or regarding whether they have been treated fairly. However, there is much psychological evidence that humans have systematic, implicit biases in judgment. Use of human arbiters thus suffers from the same difficulties as descriptive decision theory in economics: the mistakes people make are not always noisy, but are sometimes the result of systematic biases. This also undermines the ability of IF to be sufficient guard against unfairness.

3. **Prior moral judgments**: Determining whether individuals are relevantly similar requires first determining what features are relevant to fairness. Which features are task-relevant, and how they should contribute to similarity evaluation, depends on relevant moral values. Determining relevance thus requires making moral judgments about what fairness requires, prior to determining or measuring similarity. So, similar treat-

ment and IF cannot offer a substantive, non-circular definition of fairness. Moreover, IF offers inadequate guidance on its own, without appeal to other fairness judgments.

4. **Incommensurability**: Some moral values are incommensurable: they cannot be evaluated on a common measure, i.e., they cannot be straightforwardly aggregated or exchanged. This is shown by cases involving insensitivity to sweetening: decisions whose difficulty cannot be alleviated by small-value tiebreakers. If incommensurable moral values are relevant for determining similarity for some task, then this will make it impossible to represent similarity as a distance metric.

After offering additional background on individual fairness, I will discuss each of these arguments in turn.

2 Background and related work

Fair machine learning is a growing area of research that seeks to ameliorate algorithmic bias and promote algorithmic fairness. One of the chief aims of the research is to understand what fairness means in the context of algorithmic decision-making. The goal is to define fairness in a way that captures the intuitive concept with rigorous mathematical precision. Such a definition can be used as a constraint on optimizing the accuracy of ML algorithms, allowing practitioners to seek accuracy without having to worry about being unfair.

2.1 Group fairness

The dominant research approach in fair ML is known as *group fairness* (Barocas, Hardt, & Narayanan 2019). This approach attempts to define fairness in terms of statistical parity criteria (or conditions) that are imposed on the decisions or predictions produced by an algorithm. These criteria typically require that some form of statistical parity obtain between the treatment of different social groups by the algorithmic decision-maker. For instance, one such criterion requires that an algorithm produces the same false-positive rate for people from different racial groups. (This particular condition was at issue in ProPublica's famous criticism of the COMPAS risk evaluation system Angwin, Larson, Mattu, and Kirchner (2016).)

The group fairness paradigm, however, suffers from a number of problems. A wide variety of group fairness definitions have been proposed, with little agreement about which is most promising. None appears sufficient on its own to capture the intuitive notion of fairness Barocas et al. (2019); Kearns and Roth (2019). Moreover, there are counter-examples to the sufficiency of many of these criteria. These counterexamples are cases where the parity condition at issue is satisfied by the algorithm, but the decision seems "blatantly unfair" to individuals involved Dwork et al. (2012); Kearns, Neel, Roth, and Wu

(2018). Even more problematically, many of the most promising group fairness constraints are mutually incompatible: it is impossible to satisfy them at the same time Chouldechova (2017); Kleinberg, Mullainathan, and Raghavan (2016).

2.2 Individual fairness

In light of the problems for group fairness, many researchers have turned to a different paradigm, known as *individual fairness* (IF). First proposed by Dwork et al., this research program takes the essence of fairness to be that similar individuals should be treated similarly. "We capture fairness by the principle that any two individuals who are similar *with respect to a particular task* should be classified similarly (Dwork et al. 2012, p. 214). This definition has been taken up by a large body of researchers (e.g., Friedler et al. (2016); Gillen, Jung, Kearns, and Roth (2018); Ilvento (2020); Joseph, Kearns, Morgenstern, and Roth (2016); Kearns and Roth (2019); Kearns, Roth, and Wu (2017); Mukherjee, Yurochkin, Banerjee, and Sun (2020); Wang, Grgic-Hlaca, Lahoti, Gummadi, and Weller (2019)) Call this principle *similar treatment*. It is related to Aristotelian consistency principles, familiar from ethics and philosophy of law, which require that like cases be treated alike Binns (2020). Proponents of individual fairness consider similar treatment to be the intuitive definition of fairness. They seek to offer a precise mathematical treatment of the principle.

Following Dwork et al., individual fairness is typically precisely defined using two distance metrics. The first is a *similarity metric*: a distance metric that measures the degree of similarity between individuals. The second metric measures the difference in the chances two individuals have of obtaining a decision's various outcomes. Individual fairness then requires that the distance between two individuals' outcomes is no greater than their distance according to the similarity metric. Following most of the literature on IF I will appeal to the original formulation from Dwork et al. (2012).[1]

Dwork et al. use a function called a *Lipschitz mapping* to give a precise rendering of similar treatment:

Individual Fairness (IF): A mapping $M : V \to \Delta(A)$ satisfies the (D, d)-*Lipschitz* property if for every $x, y \in V$, we have:

$$D(Mx, My) \leq d(x, y) \tag{1}$$

Here, $x, y \in V$ are individuals. In most applications relevant to the fair ML literature, V is a set of individual people. M is a function that assigns to individuals a probability distribution over the outcomes A. D is distance function that measures the difference in the probabilities M assigns to two individuals. The similarity metric is represented by d.

[1] Dwork and Ilvento (2018) offer an updated version of the original Lipschitz condition. There are other variations (Friedler et al. 2016; Kearns & Roth 2019; Kearns et al. 2017), but the differences are largely irrelevant to our discussion, with a few notable exceptions to be discussed below in sections 2.4 and 3.4.

What IF says, then, is that the distance between the chances of certain outcomes assigned to two individuals (i.e., the similarity of treatment) must be no greater than the similarity-distance between the individuals (i.e., the similarity of individuals). It essentially works by measuring similarity of individuals and similarity of outcomes, representing these two measurements in a similar format (as similarly scaled distance metrics), and constraining how far apart they can be. This requires that both distance measures have real-valued output. Ensuring a real-valued output is made easier for the outcome-distance D by the fact that the outputs of M are probabilities. Dwork et al. offer several acceptable distance measures for comparing the chances of outcomes, including *statistical distance* (Dwork et al. 2012, p. 5). Our focus will be on the similarity distance metric.[2]

We can get a sense of how IF works through an example. Suppose an algorithm M is being used for college admissions decisions. M offers a "soft" prediction, giving a probability that an applicant should be admitted. Suppose two applicants, x and y are very similar: they have similar GPAs and SAT scores and come from the same high school. y is from a less wealthy family, but this fact is not considered relevant to determining who should be admitted (given certain ethical background assumptions), so the similarity metric ignores it. Thus, the distance between the applicants is very small, $d(x, y) = .01$. However, M assigns applicant x a score of .9, and applicant y a score of .7. (We can imagine it is sensitive to family wealth since it was trained on historical admissions data). According to the statistical distance metric suggested by Dwork et al., $D(Mx, My) = .2$ in this case. Thus, M would be considered (individually) unfair: it fails to respect the Lipschitz constraint. Applicant x and y are treated dissimilarly, despite being similar.

This example illustrates the intuitive appeal of individual fairness. Irrelevant differences between people should not lead to significant differences in their chance of a good outcome. There is no good justification for treating the two candidates differently. Similar treatment is meant to explain this intuitive judgment, and IF is meant to give a mathematically precise rendition of that principle.

The example also shows a feature of IF that is important for understanding some of the problems it faces: what counts as a *task-relevant* similarity or difference depends on a moral judgment about fairness. In the example, I suggested that family wealth was not relevant to determining who should be admitted. This claim depends on a moral judgment that it would be *unfair* to consider family wealth in determining who is admitted. This is an essential feature, as IF is very explicitly designed to promote "fairness through awareness" (the ti-

[2]Note that this is required to be a *metric* in the precise mathematical sense by most IF research (e.g., Gillen et al. (2018); Ilvento (2020)). Dwork et al. (2012) always refer to d as a metric, and appeal to the triangle inequality condition in their proofs. However, they suggest in a footnote that d does not need to be a full similarity metric for their purposes. Subsequent work has continued to rely on the similarity metric being a metric, e.g., Dwork, Ilvento, Rothblum, and Sur (2020); Gillen et al. (2018); Ilvento (2020). Bechavod, Jung, and Wu (2020) provide a notable exception, as they appeal only to a real-valued function. However, the four problems raised below apply also to non-metric distance functions with real-valued output.

tle of Dwork et al.'s initializing paper Dwork et al. (2012)). This means IF does *not* require the similarity metric to ignore whether individuals are members of protected groups. However, the only way to ensure that being aware of protected differences leads to fairness is for considerations of fairness to directly inform the similarity metric. I will return to this point below (section 3.3).

2.3 Arguments for individual fairness

Proponents offer two main arguments for individual fairness as a theory of fairness. The first argument concerns the intuitive appeal of individual fairness and similar treatment. As already discussed, proponents suggest that the intuitive cases show that similar treatment is the right definition of fairness, and that IF captures this principle in precise mathematical terms. In support of this, they also argue that group fairness is misguided precisely because it concerns protections for groups rather than for individuals. Fairness just is about how individuals are treated, they suggest, and so group fairness takes the wrong target of consideration. Kearns and Roth make this explicit: "... both statistical parity and equality of false negatives are providing protections for groups (in this case the two races), but not for specific individuals in those groups..." (Kearns & Roth 2019).

The second argument offered for individual fairness is that it forbids a variety of common discriminatory practices. Dwork et al. suggest that imposing the Lipshitz condition on algorithms will forbid actions such as blatant explicit discrimination, implicit discrimination (using redundant encoding), housing redlining and reverse redlining, and tokenism (Dwork et al. 2012, p. 22). More importantly, proponents argue that IF can successfully detect and prevent types of discrimination that various group fairness criteria miss. These include cases of "cherry-picking", where members of sensitive groups are randomly chosen, or are selected in a malicious way in order to undermine members of that group (Dwork et al. 2012, p. 7–8). For instance, consider a college admissions task where statistical parity between racial groups is ensured by carefully vetting the majority group applicants, while randomly selecting a proportionate number of applicants from a minority group. This seems intuitively unfair (particularly to hard-working members of the minority group who are not admitted). However, it would be compatible with a variety of group fairness criteria. IF, by contrast, forbids this kind of cherry-picking.

In what follows I raise four problems for individual fairness. These problems support my two conclusions: that IF cannot serve as a definition of fairness, and that IF should not entirely replace other forms of fairness evaluation. The four problems undermine the strength of the arguments in favor of IF, while also raising worries for the breadth of its applicability. However, note that I will not argue that the methods developed by proponents of IF should be abandoned. Ultimately, I think that these methods are valuable as one kind of tool among many needed to promote fairness and justice in machine learning. But I argue that IF methods have a more limited scope of application than its proponents suggest, and should not be seen as a complete replacement for

other fairness methods.

2.4 Related individual fairness research

There are a few alternative lines of research that do not appeal to similarity metrics but fall within the individual fairness research paradigm, broadly construed. Each of these programs has distinct issues that must be addressed. Joseph et al. (2016) and Kearns et al. (2017) offer a version of individual fairness that focuses on ensuring that individuals with greater merit always do better than those with less merit. However, this method has a serious drawback. Specifically, the distribution of merit in actual societies is affected by existing oppressive structures in those societies. A method that focuses on merit will thereby be likely to reproduce the kinds of biases algorithmic fairness is meant to ameliorate. In any case, merit-based IF differs significantly from the kind of similarity metric-based IF that is the main focus of this paper.

Yurochkin, Bower, and Sun (2020) offer a promising IF method that requires algorithms be insensitive to perturbation along protected dimensions. Their method requires that algorithms provide the same results for an individual even if their sensitive features are changed. In other words, it should give the same results to two individuals with identical features except for differences in protected class, e.g., race, gender, etc. However, this method gives up on the benefits of the "awareness" aspect of individual fairness. As noted, one of the significant benefits suggested for IF is that it allows "fairness through awareness" (Dwork et al. 2012). That the IF similarity metric is aware of an individual's sensitive features is necessary to handle cases involving corrective fairness goals (as in the *affirmative action* case below in section 3.3). Requiring insensitivity to perturbation makes it impossible to count such cases as fair, and so gives up on one of the main benefits of awareness. Weighing that drawback against the other benefits of their method is beyond the scope of this paper.

2.4.1 Related criticisms of IF

B. Johnson and Jordan (2018) formalize Aristotelian consistency, which they call the *Like Cases Maxim* (LCM), using a distance metric to characterize likeness. This formalization of LCM is highly analogous to similarity metrics for IF. They argue that the application of LCM to individual legal cases has pernicious results. For instance, LCM forbids sharp cutoffs like important age restrictions (e.g., driving age). Johnson and Jordan also note that plausible weakenings of LCM, ones that don't use similarity metrics, make the principle too weak to be helpful in promoting justice. Given how similar this formalization is to IF, Johnson and Jordan's cases provide complementary worries to the ones I raise here for IF.

3 Problems for Individual Fairness

In this section I present the four problems for individual fairness introduced above. Each gives reason to doubt that individual fairness provides a definition of fairness that captures the intuitive notion. Moreover, they offer reason to doubt that IF methods should be deployed without also appealing to other algorithmic fairness methods. For each problem, I will also discuss potential responses.

3.1 Insufficiency of Similar Treatment

The first issue I will raise for individual fairness concerns its adequacy as a definition of fairness. In particular, I will argue for the insufficiency of similar treatment for ensuring fairness. This undermines the intuitive case for IF, insofar as it relies on the idea that similar treatment is the notion which "captures" fairness.

The argument against the sufficiency of similar treatment is straightforward. If similar treatment is the right definition of fairness, then no cases in which similar treatment is satisfied should be unfair. However, it is easy to generate counterexamples to this claim. There are many cases in which similar individuals are treated similarly, but where the outcome is clearly unfair. The following two cases illustrate this:

Universal Rejection Consider a system that offers advice on college admissions decisions. In this case, the system simply recommends denying every application. Here, similar individuals are treated similarly, because everyone is treated similarly: everyone is denied admission. Despite this, individuals capable of succeeding in college, but who are denied the opportunity, can rightly complain that they have been treated unfairly.

High Risk Consider a system that assigns scores to individuals designed to measure their risk of recidivism (much like the notorious COMPAS system (Angwin et al. 2016)). This system is designed to satisfy IF, so that individuals who are relevantly similar are given similar risk scores. However, after setting it up, assume its creators adjust the system so that everyone is given the same, significant increase in risk score. The adjusted system treats everyone as a much higher risk than before it was adjusted. Here again, every similar individual is treated similarly, but everyone is treated as riskier without any justification. They may all rightly complain they were treated unfairly.

These examples illustrate the recipe for creating counterexamples to the sufficiency of similar treatment. We can generate counterexamples using arbitrary changes applied to each individual considered by the algorithm. What is required is a change that is applied equally to all individuals, and one which involves worsening the outcomes for each of them. Crucially, this will involve

an arbitrary, unjustified change, one that it would seem reasonable for each individual to reject as unfair.

In cases created with this recipe, similar treatment is satisfied, but the situation is unfair. Thus, similar treatment cannot capture the intuitive essence of fairness: it is not the definition of fairness. This undermines the intuitive argument for individual fairness as the primary way of evaluating fairness. The fact that IF is a formal characterization of similar treatment is not enough to show that IF is the right formal definition of fairness, one that is superior to other fairness criteria. The fact that IF encodes similar treatment does not show that IF is better than, or should have pride of place over, other proposed fairness criteria, as similar treatment is not the definition of fairness. At best, similar treatment is a necessary condition for fairness. This is not enough to show that IF should displace or supersede other fairness criteria.

3.1.1 Responses to insufficiency

I do not know of any extant works raising these counterexamples to the sufficiency of IF, or attempting to respond to them. However, similar treatment is a type of Aristotelian consistency principle (Binns 2020). It has been noted before that such principles are at best necessary conditions for justice, a closely related moral concept (see, e.g., Frankena (1966); Hart, Hart, and Green (2012); B. Johnson and Jordan (2018); Rawls (1971); Schauer (2018)).

One potential response to the insufficiency objection raised by the counterexamples involves distinguishing fairness from other moral concepts such as justice or goodness. Then, one may suggest that the examples show violations or deficiencies concerning one of these other moral concepts. This would explain our intuitive judgments that the cases are impermissible/deficient, without requiring that they are unfair.[3] However, I do not think this response is compelling, for several reasons. First, the most obvious alternative moral concept is justice, but it is unclear whether justice can be decoupled from fairness in the relevant way. On one of the most influential views, justice just is fairness (Rawls 2001). Second, fairness typically involves how people are treated by other agents. On many accounts, it requires that a person's treatment is adequately justifiable, or that no one could reasonably reject such treatment (Scanlon 2000). If that is correct, then fairness is clearly a relevant moral consideration in the counterexample cases, and one that is violated in them. Third, my intuition is that the treatment in these two cases is unfair. If one of the rejected candidates in *universal rejection* complained specifically that their treatment was unfair, that complaint would seem apt; we wouldn't be tempted to correct them. Finally, I will note that if we adopt this response to defend IF as a definition of fairness, the resulting notion of fairness would seem less relevant to the goal of making machine learning more ethical. After all, it shows that IF permits decisions that are clearly morally impermissible. This undermines the idea that IF should be given pride of place over other methods

[3] Thanks to two anonymous referees for suggesting this response.

for evaluating and improving algorithms with respect to their moral value or permissibility.

3.2 Systematic bias and arbiters

One of the primary issues that has stymied research into individual fairness is the difficulty of finding appropriate similarity metrics (Binns 2020; Ilvento 2020; Mukherjee et al. 2020). This is more difficult than proponents of IF initially realized (partially for the reasons discussed in section 3.3). However, recently there have been new attempts at using machine learning to derive appropriate metrics. One of the most promising directions for learning metrics involves appeal to human decision-makers as experts or "arbiters" in making similarity and fairness judgments. Ilvento (2020) appeals to human arbiters in evaluating whether individuals are relevantly similar, and builds a similarity metric based on those judgments. Similarly, Mukherjee et al. appeal to two methods for metric learning using data consisting of "human feedback, hand-picked groups of similar training examples, hand-crafted examples that should be treated similarly as observed training examples, or a combination of the above," (Mukherjee et al. 2020, p. 3). Wang et al. (2019), Lahoti, Gummadi, and Weikum (2019), and Gillen et al. (2018) offer similar attempts to learn metrics from data.[4]

Using judgments of human arbiters is a promising idea. It is commonplace in moral philosophy to appeal to people's judgments or intuitions regarding particular cases. These intuitions are typically treated as defeasible evidence that must be accommodated or explained away by a moral theory. The idea is that human judges will be reliable at recognizing similarity, fairness, and unfairness, even if they are not capable of articulating a theory of fairness. Ilvento (2020), Gillen et al. (2018), and Mukherjee et al. (2020) all explicitly appeal to similar ideas of "knowing it when you see it." Moreover, there is a long history (particularly in virtue ethics) of appealing to virtuous agents as exemplars to help guide action. It is highly plausible that human judges will be sensitive to the moral considerations that must be respected in determining what makes two individuals similar for the purpose of a particular task. Thus, use of human arbiters offers a good chance of reflecting the prior moral judgments needed for determining a fair similarity metric.

Despite the apparent promise of the project, the appeal to human arbiters to learn a similarity metric suffers from a difficulty stemming from human biases. It is well-known that humans exhibit pernicious, discriminatory biases in their judgments. Moreover, these biases need not be explicit. A large body of psychological research collected over the past five decades provides significant evidence that human judgment and decision-making suffer from systematic biases that individuals are not aware of (Brownstein 2019; Gawronski &

[4]Bechavod et al. (2020) also appeal to human arbiters (in their parlance "auditors") in pursuit of individual fairness. Unlike other attempts, however, they do not require that the similarity measure be a distance metric. While this seems promising, it does not escape the issue of systematic bias I will raise for attempts to learn similarity metrics.

Brannon 2019; Greenwald & Krieger 2006; Kahneman 2011). Much of this bias concerns rational belief and decision-making quite generally, e.g., failures to respect basic principles of probability and rational choice (Tversky & Kahneman 1974). The evidence of these biases have raised significant difficulties for using classical decision-theory for descriptive purposes in economics. And unfortunately, these implicit and systematic biases are not limited to prudential rationality. Implicit bias is a significant factor in perpetuating oppressive structures involving race, gender, and other sensitive categories.[5]

Proponents of the human arbiter approach appeal to the idea that aggregating the judgments of a large enough group of arbiters would be sufficient to limit error. This makes sense on the assumption that there is an underlying fact of the matter about similarity, and that human judgments are noisy proxies for this underlying similarity relation. However, if human judgments are *systematically* biased, rather than merely noisy, then this bias won't be washed out by aggregating larger sets of judgments. This problem is essentially the same issue that the heuristics and biases literature raised for classical microeconomics (Kahneman 2011). Attempts to learn a similarity metric from biased human judgments runs the risk of exacerbating algorithmic bias, rather than ameliorating it. That human bias might infect the similarity metric is particularly problematic since one of the main goals of fair machine learning is to help ameliorate bias.

Human judgment is unavoidable when engaging in ethical theorizing. Our intuitive moral judgments invariably and inescapably serve as part of our evidence for ethical theories. The problem of systematic bias faced by attempts to learn fair similarity metrics are thus not novel to the IF research program. Appeal to intuition in normative ethics suffers from the same worries. In that field, philosophers typically use a variety of evidence to evaluate and adjudicate intuitive judgments, including how well the judgments cohere with accepted moral principles, and the best available moral theories. This method is called *reflective equilibrium*.[6]

The crucial point about systematic bias is that it should make us wary of treating individual fairness as the sole or primary arbiter of fairness. If proponents are successful in developing methods for learning similarity metrics from human judgment for particular tasks, we should ensure that any use of an IF criteria built from such metrics is paired with other fairness evaluation

[5]There is much debate about the extent and importance of implicit bias, but recent metastudies suggest that it is a well-confirmed effect (Brownstein 2019). Some philosophers and psychologists have argued that the data meant to support the idea of implicit bias is explainable in terms of undetected *explicit biases*, perhaps involving test subjects refusal to admit explicit biases. However, this explanation offers no solace to the proponent of IF. Secret explicit biases would be just as effective at imparting pernicious discrimination into a similarity metric as genuinely implicit biases.

[6]Reflective equilibrium was originally due to Goodman (1983), but was named and popularized by Rawls (1971). As a theory about philosophical methodology, reflective equilibrium has wide but certainly not universal acceptance (Daniels 2020). Detractors, however, typically appeal to some alternative way of evaluating intuitive judgments. Few advocate completely ignoring such judgments, and even fewer advocate complete fidelity to them.

methods. One clear benefit of group fairness statistical criteria is that they will be sensitive to algorithmic biases that emerge only in the aggregate, ones which may be inherited unnoticed by similarity metrics used for individual fairness criteria.

3.2.1 Responses to systematic bias problems

Researchers working on learning similarity metrics are appropriately concerned with the difficulty of building a similarity metric for fairness. They appeal to human arbiters given the fact that human judgments are essential as evidence and guidance regarding fairness. These IF researchers are sensitive to the worry that humans might be unable to accurately answer the kinds of difficult questions being asked them (for instance, see (Ilvento 2020, Sec. 1.7)). They are also sensitive to worries about moral disagreement (e.g., see (Mukherjee et al. 2020, Sec. 2)). To my knowledge, however, there is no explicit consideration of worries about implicit bias. The best candidate for help with correcting for learned implicit bias would be to pair IF methods with other fairness criteria, like the group fairness statistical parity criteria, as in Dwork et al. (2012) and Zemel, Wu, Swersky, Pitassi, and Dwork (2013). Group fairness criteria will be less prone to implicit bias than learned metrics (though not immune). Using the two kinds of methods together may offer improved protection by increasing the chance of catching the bias.

3.3 Prior moral judgments

The third problem I will raise for individual fairness also concerns its adequacy as a definition of fairness. Moreover, it suggests that finding an accurate similarity metric may be even more difficult than has been understood. The problem stems from the fact that the similarity metric needed for IF depends on prior fairness judgments, i.e., moral judgments concerning what it is fair to consider task-relevant.

One of the arguments for individual fairness is that it rules out certain kinds of unfair cases that group fairness criteria permit. In order for IF to work this way, or for any purpose, it requires the practitioner to know the similarity metric d. That is, we must know a function that assigns to each pair of individuals a distance between them, represented by a real number. In order for IF to accurately represent fairness for a task, by the lights of IF's proponents, this similarity metric must accurately reflect the task-relevant similarities and differences among individuals.

As noted above, the similarity metric must be sensitive to moral features of the task and of the individuals, in addition to the relevant descriptive features. This fact was recognized by Dwork et al. (2012, p. 3), who suggest that metrics must be adjusted in a way agreed upon by members of society. Their appeal to political agreement to justify claims about fairness and justice is inspired by Rawls (2001).

A pair of examples will help show the importance of moral judgments about fairness to building an adequate similarity metric. I will continue appealing to college admissions as a useful kind of task for our examples.

Discriminatory Admissions A system M is used to assist decisions concerning admissions for a university. Its task is to select successful students: those who are likely to graduate, obtain high GPAs, promote university reputation via extra-curricular activities, and obtain subsequent employment. M is trained on historical data from the university's admissions committee's decisions, learning to mimic past human decisions. In the data, white applicants were admitted at higher rates than black applicants. M thus learns a preference for white applicants and admits them at higher rates even where other features are equivalent. This turns out to be highly accurate for the task of choosing successful students as the university environment and subsequent job market are filled with implicit and explicit racism that negatively impact black students' success.

Affirmative Action M is being used for a different university. As before, it is used to assist admissions decisions with a goal of selecting students who will be successful. It is again trained on historical data. This university aims to promote diversity and uses various methods of affirmative action. For instance, it adds some points to black applicants' SAT scores, in light of potential bias in the test itself, and the fact that students from this group typically have less access to test preparation. Thus, M learns to admit more black applicants for this university.

I think it is clear that these two cases are morally very different. Even if you are unconvinced that affirmative action is permissible, the second case is clearly better than the discriminatory admissions case, which involves continuing and exacerbating the oppression of a marginalized group. However, both cases can be interpreted as "treating similar people similarly." Both involve taking into consideration racial differences that are relevant to the success of M at achieving its task. Moreover, in both cases, one can construct a similarity metric d that would satisfy the Lipschitz condition and (formally) satisfy individual fairness. In the former case, race is used as a reliable predictor of student success. In the latter, it is used to promote diversity in the student body (and potentially redress historical injustice). In both cases, race is a feature that d is sensitive to, a feature that makes for greater differences between individuals. Intuitively, one of these uses of race as a reason for dissimilar treatment is unfair, and one of them is fair (or at least less unfair).

This illustrates a crucial feature of individual fairness: it requires antecedent moral judgments about which features of individuals are *fair* to consider task-relevant. Building the similarity metric d requires moral judgments about what kinds of differences are fair to treat as task relevant, and which ones are not. Moreover, such judgments are necessary for determining *how* task-relevant features may fairly contribute to our evaluations of similarity and dissimilarity. The only way to distinguish between the acceptable use of race as a

feature in *affirmative action* and the morally unacceptable use in *discriminatory admissions* is to make a substantive moral judgment about fairness. In sum, *treating similar individuals similarly requires appeal to substantive moral claims about who it is fair to count as similar or dissimilar*. Neither similar treatment nor its precisification into individual fairness offers guidance on how to make such fairness judgments.

Moreover, this feature of individual fairness is not a minor side note: it is crucial to the way IF promotes "fairness through awareness." Using sensitive or protected features like race, gender, lgbtq+ status, etc. is essential to how IF is supposed to deliver fairness. It's well-known that attempts to achieve fairness through unawareness do not work in a wide variety of cases (Barocas et al. 2019). When systems are designed that do not have access to the sensitive features, they will still reproduce bias through a number of means, including using proxies for the sensitive features (G. M. Johnson 2020). One of IF's chief virtues is the way it is aimed at solving this problem.

Our *affirmative action* case shows how prior fairness judgments are necessary for IF to help promote fairness through awareness. An appropriate similarity metric, d, for the task in *affirmative action* will treat a white applicant x with an SAT of 1400 as relevantly similar to a black applicant y with a score of 1250. In other words, $d(x, y)$ will be small. The similarity metric treats individuals with these differences as similar. Building the right similarity metric here requires making the judgment that this kind of sensitivity to racial categories is fair. Making such moral judgments in building d is a necessary feature for allowing IF to promote fairness through awareness. In contrast, the kind of sensitivity to social categories shown in *discriminatory admissions* is unfair. But nothing about the concept of similarity, by itself, does the work of distinguishing the two cases.

The necessity of making fairness judgments in building a similarity metric is another factor that undermines the intuitive argument for individual fairness. Similar treatment is purported to be the right definition of fairness, one that offers guidance on achieving fairness. IF is supposed to be preferable to group fairness notions because it encodes similar treatment, a principle which is supposed to capture the essence of the pre-theoretic, intuitive notion of fairness. Similar treatment and IF are purported to be informative, non-circular definitions of fairness. However, similar treatment and IF cannot deliver on this promise of offering guidance, or on avoiding circularity. This is because they rely on antecedent moral judgments about what features it would be *fair* to treat as relevant to determining similarity. Instead of giving us guidance on fairness for a task, IF relies on our having prior knowledge of what is fair. Thus, IF does not provide a substantive, non-circular definition of fairness.

For the reasons noted in the section 3.1, similar treatment offers at most a necessary condition on fairness. Necessary conditions, however, are not always informative. Being born on Earth is (at least at the moment) a necessary condition on being accepted to college, but this does not tell us very much about who will be admitted. Similar treatment and IF are meant to provide guidance on what fairness is and what it requires. If we don't know what the

fair thing to do is, IF purports to tell us. As we have seen, however, it doesn't offer as much independent guidance as it is supposed to. Instead, it requires prior moral judgments about what features are fair to consider for a task. The required similarity metric cannot be constructed before we know this.

To be clear, this reliance on prior fairness judgments does not show that similar treatment and IF are false as principles. Nor does it suggest that we should abandon individual fairness as a method for assessing fairness. IF might still represent a useful way of aggregating information about fairness judgments. What it does undermine is the idea that IF is *the* definition of fairness, one with pride of place over other kinds of fairness criteria in virtue of its derivation from the fundamental or essential nature of fairness. Similar treatment, as we have seen, is at best a necessary principle of fairness. Moreover, it is one that depends for its content on antecedent fairness judgments. For these reasons, it cannot be *the* definition of fairness. Nor can IF serve as the singular or primary method for assessing fairness.

The problem presented here is not novel to individual fairness in machine learning. Similar treatment is a type of Aristotelian consistency principle, as Binns (2020) notes. Consistency principles of this sort have been appealed to in both philosophical ethics and the law. Typically, they are expressed as the requirement that "like cases be treated alike." For such requirements, the same kind of issue has arisen: they offer little in the way of the guidance needed to determine when two situations are alike. This has been called the problem of *emptiness*. Aristotelian consistency principles are "merely formal": they require substantive moral commitments about what it means for two cases to be relevantly alike. They are empty from the perspective of informativeness and action guidance. In contemporary philosophical ethics, this issue has been noticed since at least (Frankena 1966). Schauer (2018) makes a similar point regarding consistency principles that govern judges' decisions in legal cases. It might be true that cases which involve the same issues should be given the same verdicts; but what it means for two cases to be the same is precisely what is at issue in most disputes on the subject. The principle thus offers little information or guidance.

The issues I have raised here for individual fairness involve the same deep issue for consistency principles. In addition to undermining the intuitive case for treating IF as the pre-eminent definition of fairness, this helps explain the difficulties researchers have faced in building similarity metrics. Knowing that one is looking to account for similarities is not enough to develop such a metric; one must also be sensitive to particular moral judgments about fairness. This also suggests that similarity metrics developed for other purposes will often not be suitable for use in individual fairness evaluation. For instance, Dwork et al. (2012, p. 3) suggest that a similarity metric could be found from a system for diagnosing cardiology patients, the Advanced Analytics for Information Management project.[7] However, this project is unlikely to include the moral

[7]For details about this project, see https://web.archive.org/web/20120209000934/http://www.almaden.ibm.com/cs/projects/aalim/

fairness judgments required for building the kind of similarity metric needed for individual fairness.

3.3.1 Responses to prior moral judgment worries

Dwork et al. (2012) suggest that, in the absence of knowledge of the real similarity metric, a metric can be constructed using "the 'best' available approximation as agreed upon by society" (p. 214). They cite Rawls (2001) as inspiration. This suggests a recognition that moral judgments must be made in order to build a similarity metric. Research into learning fairness metrics, such as work by Gillen et al. (2018), Ilvento (2020), and Mukherjee et al. (2020) (discussed above in section 3.2) also shows awareness of the importance of prior fairness judgments in building similarity metrics. However, these works have not given arguments defending the non-circularity and informativeness of similar treatment and IF in light of their dependence on prior moral judgments.

A potential response would be to highlight the usefulness of the similarity metric as an aggregation of the prior moral judgments about fairness. This aggregation response concedes that the prior moral judgments are explanatorily prior to IF. It thus does not defend IF as a substantive, non-circular definition of fairness. However, it does offer a role for IF, and so it defends against the objection that IF is unhelpful or empty of advice. On this view, IF does offer advice about avoiding unfairness. While it's true that doing so requires appeal to prior fairness judgments, once those judgments are made the similarity metric helpfully aggregates them, and the Lipschitz condition helps identify cases of unfairness. This response seems promising, and fits well with the research program into learning similarity metrics from human arbiters discussed in the last paragraph and in section 3.2. However, it will only work for tasks where there is a coherent metric to be built. In section 3.4, I discuss why, for many tasks, there will fail to be a similarity metric.

One might expand on this aggregation response to give a defense of similar treatment and IF as definitions of fairness by appealing to the method of reflective equilibrium (mentioned above in section 3.2).[8] This method involves bringing our intuitive moral principles and intuitive moral judgments into coherent alignment, a process which results in a coherentist justification for our reflective, considered principles and judgments (Daniels 2020). The current response suggests that accepting IF brings our moral principles and our intuitive judgments into coherent alignment, and so we are justified in believing IF. While this response does speak to the justification of IF, it does not show that IF offers a substantive, non-circular definition of fairness. Being *justified* as believable or acceptable does not show that a principle provides an informative definition of a concept. Most of our true and justified beliefs are not definitional of anything. Thus, this response does not undermine the point that IF is not an informative, non-circular definition of fairness. This point is all that is

[8] Thanks to an anonymous referee for suggesting this response.

needed to show that IF should not be given price of place over other methods in virtue of its uniquely capturing or providing the definition of fairness.

3.4 Incommensurability

As I have argued, an accurate similarity metric requires prior moral judgments about what kinds of similarities are task-relevant. These moral judgments must accurately track moral values.[9] This raises another worry for developing accurate similarity metrics. The moral values in question must be commensurable in order for them to be combined into a similarity metric. That is, the existence of a similarity metric requires that it be possible to aggregate the moral values, or evaluate them together, in a straightforward way. Crucially, this straightforward aggregation must allow for tradeoffs between the values. However, philosophers have argued that a variety of moral values are *incommensurable* with one another, meaning they cannot be aggregated in terms of a common measure (Chang 2015; Hsieh 2016). If the values relevant for a particular task are incommensurable, it will be impossible to construct a real-valued similarity metric as used by proponents of individual fairness.

Incommensurability is a relation between two (or more) values. Two values are incommensurable iff there is no common measure that can be applied to both values. Basically, values are incommensurable if there is no way of exchanging or trading them off in a predictable, straightforward manner. This point was historically raised as a problem for treating money as the common measure of all value (Chang 2013). However, the same point will apply to other measures, particularly those which can be represented by real-values like similarity metrics.

One way for values to be incommensurable is as a result of incomparability. Incomparability is a relation that holds between *bearers of value*, i.e., things that have value. Bearers of value include actions, choices, outcomes, or individuals. The sense of "comparable" relevant here means *able to be compared*.[10] Two items are comparable if they bear a comparative relation to one another. According to the traditional view, known as the *trichotomy thesis*, there are exactly three comparative relations: *better than*, *worse than*, and *equally good* (Chang 2002). For instance, if you prefer the taste of chocolate to the taste of prunes, then the action *eating chocolate* is better than *eating prunes* in terms of

[9] All of the points in this paper are compatible with a variety of metaethical views about the nature of moral judgments and moral values (Sayre-McCord 2014). For instance, even expressivists will agree that there can be inaccurate moral judgments: those that do not actually express the emotions, intentions, or plans of the speaker. Moreover, one need not be committed to moral realism about the relevant values here to recognize the incommensurability worries I raise in this section, as incommensurabilty and parity cases arise even for individual preferences in decision-making (Chang 2005; Hsieh 2016).

[10] Sometimes "comparable" is used to mean *similar* by individual fairness researchers (Yurochkin et al. 2020). That is not what I mean here. There is also some dispute in the philosophical literature regarding the terms "incommensurable" and "incomparable". I follow the usage suggested by Chang (2015) and Hsieh (2016). I focus on incommensurability that arises as a result of either incomparability or Chang's notion of parity (2002).

your gustatory values. Thus, the two actions are comparable. The trichotomy thesis entails that the only way for a rational agent (with full knowledge of her options) to have no preference between two comparable choices is if the choices are equally good. If a person knows all about the relative tastiness of chocolate and prunes, yet she has no preference between the two, then the two choices must be equally good (in terms of taste) to her.

Traditional decision theory assumes the trichotomy thesis, and assumes that all of an agent's choices are comparable (Schoenfield 2014). These assumptions are necessary for representing an agent's values using a utility function. Utility functions assign a real number value to outcomes, representing how much the agent values that outcome. In a utility function, the *better than* relation is represented by $>$, *worse than* by $<$, and equally good by $=$. Every pair of real numbers x, y is related by exactly one of the relations $\{>, <, =\}$. Moreover, if it's not the case that $x > y$ and it's not the case that $y > x$, then $x = y$. As a result, if an agent's values are accurately represented by the real numbers that are output by a utility function, the only way for the agent to have no preference between two options is for them to have equal value. In other words, the agent must view them as equally good. If there is at least one pair of options to which none of the three relations apply, then there can be no utility function describing the agent. The similarity metrics used for IF also involve a function with a real number as output. A similar issue arises for similarity metrics, for similar reasons, to be discussed shortly.

To illustrate the problem of incommensurable moral values we can consider cases of *insensitivity to sweetening* (Chang 2002; De Sousa 1974; Hsieh 2016). In such cases, an agent has difficulty choosing between two options, and small improvements in either choice do not seem to help them make up their mind.

Hard Career Choice After finishing her Ph.D, Adele must choose between option A, a job for a non-profit charity, or option B, a job as an assistant professor. Working as an academic would allow her to do research and teach, both of which she thinks are morally valuable. Working at the charity would allow her to alleviate people's suffering and enact solidarity with others, both of which she again thinks are valuable. Adele finds herself having difficulty choosing. Both options concern moral values she cares about, and she is uncertain which choice would be better. Finally, in order to help break the tie, Joseph offers Adele twenty dollars if she takes the professor job.

What this case illustrates is that hard choices, in which the agent has no clear preference between the two options, are not always cases where the options are *equally good*. Intuitively, it would be perfectly reasonable for Adele to remain undecided between the options, even after one of them has been "sweetened" with a small improvement. Cases like this one are common, so even if this case does not elicit your intuition that Adele is rational, it is likely there are other cases of hard choices that do seem rational.

Cases of insensitivity to sweetening, like *hard career choice*, show that even if a rational agent does not have a preference between options A and B, it is

sometimes still rational for her to have no preference between $A + \$20$ and B. This is true even assuming she prefers $A + \$20$ to A. For these preferences to be rational, either A and B must be incomparable, or the trichotomy thesis must be false. Either way, some of Adele's values must be incommensurable. There will be no way to build a utility function that accurately represents her values in this case.

Another way to see the problem posed by this case involves the condition of *negative transitivity*. This condition requires that for any three options O_1, O_2, O_3, if an agent does not prefer O_1 to O_2, and they do not prefer O_2 to O_3, then they must not prefer O_1 to O_3. An agent must satisfy negative transitivity in order for there to be a utility function that represents her preferences. This is again because utility functions use real numbers to represent values, which necessarily satisfy negative transitivity for $\{<, >, =\}$. Adele's preferences in this case fail to exhibit negative transitivity. She does not prefer $A + \$20$ to B, she does not prefer B and A, but she *does* prefer $A + \$20$ to A. So there can be no utility function that represents her.

Insensitivity to sweetening cases show that either the trichotomy thesis is false or that the choices are incomparable. Chang (2002) argues that we should reject the trichotomy thesis, and instead recognize a fourth relation she calls *parity*. Broome (1997), in contrast, argues that we can maintain the trichotomy thesis by recognizing that comparability is a vague concept, and that sweetening cases involve indeterminately comparable values. The results of this dispute won't concern us. The existence of either incomparability, parity, or indeterminacy between choices in these cases is sufficient for incommensurability between the values in question.

There are insensitivity to sweetening cases which pose direct problems for building a similarity metric for individual fairness applications. Consider a case involving college admissions once again.

Hard Admissions Choice The admissions committee for a university must decide between two candidates for admission, Bridget and Claire. The committee is seeking to promote moral values such as student success, intellectual achievement, a nurturing community, diversity, etc. Bridget has a 3.8 GPA, a 1300 SAT score, and a history of volunteering for charity. Claire has a 3.8 GPA, a 1300 SAT score, and won several student science competitions. The committee finds the choice difficult to make, and is unsure which applicant would be better to admit in order to promote their values. They end their first meeting undecided. At the next meeting, they learn Claire has taken the SAT again, and this time received a score of 1320.

In this case, it seems reasonable for the committee not to treat the extra 20 points of the SAT score as a tie-breaker. It would be rational (and morally permissible) for them to still find it hard to choose between the two applicants, even after the additional "sweetening" of the higher SAT score.

If the committee remains unconvinced by the extra 20 SAT points, then the options *B admitting Bridget* and *C admitting Claire* could not have been equally

good to begin with. Moreover, the committee's valuation will fail to be negatively transitive. Where $C+$ is the option of admitting Claire with an additional 20 points on her SAT, the committee does not prefer $C+$ to B, and they do not prefer B to C, but they do prefer $C+$ to C. Therefore, at least some of the values at issue in *hard admissions choice* — namely, student success, intellectual achievement, community, and diversity — must be incommensurable.

The incommensurability of the values in *hard admissions choice* is a problem for building a similarity metric for individually fair college admissions. As argued above (section 3.3), similarity metrics require prior moral judgments regarding what makes for task-relevant similarity. Accurate moral judgments must track the underlying values in the case. A similarity metric for the applicants must aggregate the values of student success, intellectual achievement, and community. Thus, building a similarity metric for this case requires moral judgment about fairly aggregating these values. For instance, it must determine whether an applicant who will improve diversity and community is similar to another who will increase the community's intellectual achievement. However, if the moral values in question are incommensurable, then there can be no accurate judgment that allows for a precisely quantified, real-numbered aggregation of these values into a similarity metric (or even a non-metric, real-valued function).

Suppose we attempted to build a similarity metric that included Bridget (B), Claire (C), and Claire + 20 SAT points ($C+$). We assume there is at least one other student, Alice (A), that must also be categorized by the metric. For ease of exposition, imagine Alice is the ideal student, so distance from Alice is distance from the ideal (this assumption isn't necessary for the argument, but makes it easier to state). Since the committee cannot decide between Bridget and Claire, their distance to the ideal must be equal: $d(A,B) = d(A,C)$. Moreover, the "sweetening" of Claire's application with 20 SAT points does not sway them, so $d(A,B) = d(A,C+)$. In addition, Claire's application is clearly better after her SAT scores improve, so Claire+ is closer to the ideal student, Alice: $d(A,C+) < d(A,C)$. However, now we have arrived at the same difficulty raised above for real-valued utility functions: there is no way to assign real numbers to the set of equations/inequalities just described. If $d(A,B) = x$, $d(A,C) = y$, and $d(A,C+) = z$, then we would need an assignment that makes the following claim true: $x = z < y = x$. This is clearly impossible. Thus, if there are incommensurable values at play, it will be impossible to construct a real-valued similarity measure.

3.4.1 Responses to Incommensurability

Value incommensurability is not a problem explicitly considered in the individual fairness literature. Philosophers have attempted to provide alternative decision theories that allow for parity and incommensurability (Gert 2004; Hare 2010). However, these alternatives face significant difficulties, as they require the denial of very plausible principles regarding the relationship between expected value and actual value (Bales, Cohen, & Handfield 2014;

Schoenfield 2014). For example, they deny a principle Schoenfield calls Link: "... If you are rationally certain that neither of the two options [A or B] will bring about greater value than the other, it's not required that you choose A, and it's not required that you choose B" (2014, p. 267). Schoenfield points out that one alternative decision theory which potentially escapes this difficulty is described but rejected by Hare (2010), which he calls *deferentialism*. It is unclear how this theory would be applied to building a similarity metric, but this is a potential option for future research.

Another potential response would be to eliminate the use of a similarity metric or any real-valued function. Instead, IF could appeal to a partial-order ranking of individuals in terms of similarity. This idea is promising, and I intend to explore it in future research. However, the proposal faces immediate hurdles. For one thing, such a ranking would only be able to offer weaker constraints on treatment. A metric is a fine-grained representation that reflects the magnitude of individual differences. Rankings made with partial orderings would be much less fine-grained and would not represent magnitude of differences. In the *hard admissions choices* case, the similarity ranking would be forced to rank Bridget, Claire, and Claire + 20 points as tied. Moreover, a ranking like this would involve some distortion of the similarities. After all, Claire's application is clearly better after her SAT score improves.

Finally, one may think there is a *tu quoque* response available to the proponent of IF: won't the incommensurability of values be a problem for group fairness (GF) conditions, also? However, it is not obvious how the objection would be carried over to GF. Recall (from section 2) that GF conditions involve statistical parity requirements between protected social groups, e.g., that risk scores have equal error rates for both black and white defendants (Chouldechova 2017). GF does not require a metric which evaluates how similar people are to one another *in order to evaluate fairness*. The emphasis is to highlight that the ML methods *being* evaluated may indeed involve metrics—but these metrics need not be sensitive to incommensurable moral values. It is only because IF uses a metric *for* evaluating fairness that it has this difficulty. GF methods only require that we accurately categorize individuals' group membership. This is a genuinely difficult issue, particularly concerning socially constructed group categories such as race (Hanna, Denton, Smart, & Smith-Loud 2020), and because of intersectionality (Hoffmann 2019). But this issue is distinct from the problems raised here concerning the use of a real-valued similarity measure.

4 Discussion and Conclusion

I raised four problems for individual fairness. The problems each give support for two conclusions: first, that individual fairness does not provide a definition of fairness that captures the intuitive notion; and second, that IF methods should not be given priority over, or used in isolation from, other fairness methods. Moreover, the argument from incommensurability suggests

that there are cases in which current IF methods will be completely inapplicable. In cases involving incommensurability, building a similarity metric that accurately reflects task-relevant moral values will be impossible.

What is fair about individual fairness, then? I have not argued that IF will never be useful in promoting fairness. Plausibly, the motivating principle behind IF, similar treatment, is required for fairness in many cases. At the very least, clear violations of similar treatment provide evidence that the algorithmic predictions or decision-making in question are unfair. This suggests that individual fairness should be treated as an important tool for diagnosing unfairness—just not the unique tool, to be used instead of, or in isolation from other methods. Satisfaction of IF, at least for tasks that don't involve incommensurable moral values, provides defeasible evidence that the decision-making was fair. This evidence may be overridden in light of other evidence, e.g., unexpected violations of group fairness criteria. The idea of pairing IF with statistical parity conditions was discussed from the beginning of the research program by Dwork et al. (2012). In that paper, putting group and individual fairness together was treated as a secondary, less desirable option. However, Zemel et al. (2013) provide a more irenic proposal that treats the two methods as equally important. Along with Binns (2020), I think this idea of using the methods together should be standard. At the same time, tasks for use with IF must be chosen carefully. Moreover, methods for learning similarity measures from human arbiters must be undertaken with caution given the existence of systematic human bias.

Acknowledgments

For helpful discussion and feedback on this paper I would like to thank Tina Eliassi-Rad, Branden Fitelson, D Black, Frank Wu, Lisa Miracchi, the Fall 2020 University of Pennsylvania graduate seminar in Philosophy of AI and Robotics, and several anonymous reviewers.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May). Machine bias. *ProPublica*.

Bales, A., Cohen, D., & Handfield, T. (2014). Decision theory for agents with incomplete preferences. *Australasian Journal of Philosophy*, *92*(3), 453–470.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org.

Bechavod, Y., Jung, C., & Wu, S. Z. (2020). Metric-free individual fairness in online learning. In *Advances in neural information processing systems*.

Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 514–524).

Broome, J. (1997). Is incommensurability vagueness? In R. Chang (Ed.), *Incommensurability, incomparability, and practical reason*. Harvard University Press.

Brownstein, M. (2019). Implicit Bias. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

Chang, R. (2002). The possibility of parity. *Ethics*, *112*(4), 659–688.

Chang, R. (2005). Parity, interval value, and choice. *Ethics*, *115*(2), 331–350. doi: 10.1086/426307

Chang, R. (2013). Incommensurability (and incomparability). *International encyclopedia of ethics*.

Chang, R. (2015). Value incomparability and incommensurability. In I. Hirose & J. Olson (Eds.), *The oxford handbook of value theory*. Oxford University Press.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153-163. doi: 10.1089/big.2016.0047

Daniels, N. (2020). Reflective Equilibrium. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2020 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/.

De Sousa, R. B. (1974). The good and the true. *Mind*, *83*(332), 534–551.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

Dwork, C., & Ilvento, C. (2018). Fairness Under Composition. In *10th innovations in theoretical computer science conference (itcs 2019)* (Vol. 124, pp. 33:1–33:20). doi: 10.4230/LIPIcs.ITCS.2019.33

Dwork, C., Ilvento, C., Rothblum, G. N., & Sur, P. (2020). Abstracting fairness: Oracles, metrics, and interpretability. In *1st symposium on foundations of responsible computing* (pp. 8:1–8:16).

Frankena, W. (1966). Some beliefs about justice: Lindley lecture. *University of Kansas: Lawrence*.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*.

Gawronski, B., & Brannon, S. M. (2019). Attitudes and the implicit-explicit dualism. *The handbook of attitudes*, *1*, 158–196.

Gert, J. (2004). Value and parity. *Ethics*, *114*(3), 492–510.

Gillen, S., Jung, C., Kearns, M., & Roth, A. (2018). Online learning with an unknown fairness metric. In *Advances in neural information processing systems* (p. 2605–2614).

Goodman, N. (1983). *Fact, fiction, and forecast*. Harvard University Press.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California law review*, *94*(4), 945–967.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Facct '20* (pp. 501–512).

Hare, C. (2010). Take the sugar. *Analysis*, *70*(2), 237–247. doi: 10.1093/analys/anp174

Hart, H. L. A., Hart, H. L. A., & Green, L. (2012). *The concept of law*. Oxford University Press.

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

Hsieh, N.-h. (2016). Incommensurable Values. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2016 ed.). Metaphysics Research Lab, Stan-

ford University. https://plato.stanford.edu/archives/spr2016/entries/value-incommensurable/.

Ilvento, C. (2020). Metric learning for individual fairness. In *1st symposium on foundations of responsible computing* (pp. 2:1–2:11).

Johnson, B., & Jordan, R. (2018). *Should Like Cases Be Decided Alike?: A Formal Analysis of Four Theories of Justice* (Tech. Rep.). Social Science Research Network. doi: 10.2139/ssrn.3127737

Johnson, G. M. (2020). Algorithmic bias: on the implicit biases of social technology. *Synthese*, *in press*, 1–21.

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *Advances in neural information processing systems* (pp. 325–333).

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 2564–2572).

Kearns, M., & Roth, A. (2019). *The ethical algorithm*. Oxford University Press.

Kearns, M., Roth, A., & Wu, Z. S. (2017). Meritocratic fairness for cross-population selection. In *Proceedings of the 34th international conference on machine learning* (pp. 1828–1836).

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

Lahoti, P., Gummadi, K. P., & Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)* (p. 1334-1345). doi: 10.1109/ICDE.2019.00121

Mukherjee, D., Yurochkin, M., Banerjee, M., & Sun, Y. (2020, 13–18 Jul). Two simple ways to learn individual fairness metrics from data. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 7097–7107).

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Harvard University Press.

Sayre-McCord, G. (2014). Metaethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2014 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2014/entries/metaethics/.

Scanlon, T. (2000). *What we owe to each other*. Belknap Press.

Schauer, F. (2018). On treating unlike cases alike. *Constitutional Commentary*, *34*, 1–19.

Schoenfield, M. (2014). Decision making in the face of parity. *Philosophical Perspectives*, *28*(1), 263–277. doi: 10.1111/phpe.12044

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K. P., & Weller, A. (2019). An empirical study on learning fairness metrics for compas data with human supervision. *arXiv preprint arXiv:1910.10255*.

Yurochkin, M., Bower, A., & Sun, Y. (2020). Training individually fair ML models with sensitive subspace robustness. In *Proceedings of the 8th international conference on learning representations*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th international conference on machine learning* (pp.

325–333).