Preprint version - please cite published version:
https://link.springer.com/article/10.1007/s11158-022-09561-4

Abstract. Prediction-based decisions, which are often made by utilizing the tools of machine learning, influence nearly all facets of modern life. Ethical concerns about this widespread practice have given rise to the field of fair machine learning and a number of fairness measures, mathematically precise definitions of fairness that purport to determine whether a given prediction-based decision system is fair. Following Reuben Binns (2017), we take "fairness" in this context to be a placeholder for a variety of normative egalitarian considerations. We explore a few fairness measures to suss out their egalitarian roots and evaluate them, both as formalizations of egalitarian ideas and as assertions of what fairness demands of predictive systems. We pay special attention to a recent and popular fairness measure, counterfactual fairness, which holds that a prediction about an individual is fair if it is the same in the actual world and any counterfactual world where the individual belongs to a different demographic group (cf. Kusner et al. 2018).

1. Introduction

Prediction-based decisions—decisions based on statistical predictions—influence nearly all aspects of modern life. Consider targeted online ad delivery, which most of us encounter on a regular basis. It is commonplace for websites and applications to deliver targeted advertising, whose content for a given user is determined by some kind of prediction about them. The popularity that prediction-based decision making has recently enjoyed in online ad delivery and other applications, such as criminal justice, education, elections, employment, entertainment, finance, housing, and medicine, is due in large part to the advancements in *machine learning*—the study of computer applications that can learn from data, often in ways that can be automated or do not require explicit instructions to the machine.

The increased reliance on prediction-based decision making has been accompanied by growing concerns about the fairness of the use of this technology. These concerns are not unfounded. Machine learning programs often inherit biases from the data they are trained on, without any direct influence from their programmers (Johnson 2020). To see what this might look like on the ground, consider Latanya Sweeney's discovery in 2013 that ads by Google AdSense--an advertising service provided by Google--were more likely to show ad copy suggestive of previous arrests (e.g., "*Trevon Jones, Arrested?*" for the search "*Trevon Jones*") for searches of names assigned primarily to Black babies, independently of whether the company sponsoring the ad had a criminal record for the searched name (Sweeney 2013).

These sorts of concerns and findings have given rise to the field of fair machine learning and a number of *fairness measures*, mathematically precise definitions of fairness that purport to determine whether a given prediction-based decision system is fair. Note that when the fair machine learning community talks about fairness they often have something very specific in mind, namely that a system is fair if and only if it is not wrongfully discriminatory.[1] But this characterization of

[1] Note, then, that a "fair" machine could be fair in this sense and yet strike us as unfair in another. For instance, it might be grossly inaccurate.

(un)fairness leaves open exactly what standard to use when judging whether a decision is unfair in this way.

While many proposals have been made as to how to measure fairness in the context of prediction-based decision making, there is little agreement as to which—if any—of these proposals are correct. This is partially explained by the fact that, as Binns (2017, p. 2) has put it, "`fairness' as used in the fair machine learning community is best understood as a placeholder term for a variety of normative egalitarian considerations".[2] Binns's insight highlights, we think, the fact that the phrase "wrongfully discriminatory" typically expresses an egalitarian standard of wrongfulness, but not always the exact same standard. There are many different egalitarian standards we might care about, and which ones are operative could vary from context to context. This, in turn, will inform our choice of a fairness measure, which, we take it, operationalizes the operative standard(s) that we care about in a given context. This means—among other things—that there could be a plurality of correct fairness measures, and that the failure of a measure to capture all we care about with respect to fairness in some context is not enough to show that the measure fails to capture the relevant fairness concerns in some other context.[3]

In our view, this pluralism about fairness measures is correct. And, in the course of making our argument in this paper, we hope to offer good reasons for thinking that it is. But that is not our main aim. Rather, we are interested in exploring *how* to choose a fairness measure within a context and thinking about *when* certain measures should or shouldn't be used. In the service of these goals, we will present a general picture for thinking about the choice of a fairness measure, discuss the choiceworthiness of three measures ("fairness through unawareness", "equalized odds", and "counterfactual fairness"), and draw from our discussion some general lessons about measuring fairness.

2. A general picture

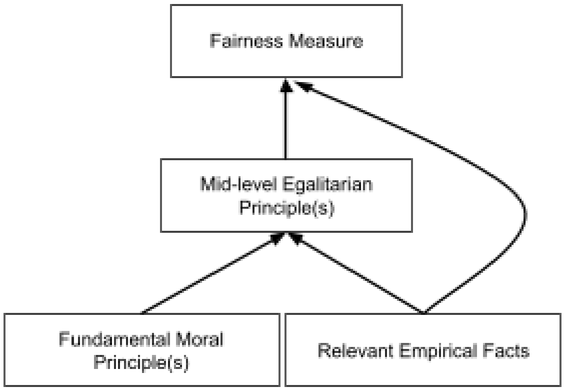
Typically, discussions of fairness measures relate a few measures to one another and draw on cases in an attempt to show that the verdicts of the respective measures are intuitively right or intuitively wrong. But these discussions rarely involve significant engagement with the normative ideas—such as the specific egalitarian ideas—that might justify the choice of one measure over another. This, as we will soon demonstrate, leaves important tools for evaluating fairness measures on the table.[4]

[2] This is no doubt also partially due to the fact that several intuitive measures are incompatible with each other in large ranges of cases, forcing hard choices between measures. For more on this, see Kleinberg et al. (2016).

[3] We also think that "fairness" does even more than this. For instance, we think "fairness" is also, at times, a placeholder for a less morally laden, statistical sense of "fair" that is synonymous with "unbiased," in the sense of lacking errors that systematically skew results towards certain types of outcomes over others.

[4] Heidari et al (2018) is a notable exception. Heidari et al argue that certain fairness measures can be given a natural interpretation as operationalizing a concern for certain egalitarian principles, and that users of fairness measures have reason to select only a measure that reflects what, in the relevant context, fairness requires. As will emerge in our discussion below, we fully agree. We take ourselves to building on this work (and that of Binns (2017)) by further investigating how robust are the justificatory connections between particular pairs of fairness measures and normative principles.

To this end, it will be helpful to sketch a general picture of the structure of the justification of a measure. If a fairness measure is appropriate in some context (say, college admissions), it is so because it tracks the verdicts of what we will call the relevant *mid-level egalitarian principles* that are implicated in that context. And which (if any) mid-level egalitarian principles are implicated in a context is a function of the relevant facts about that context and the correct *fundamental moral principles*. The general picture, then, is this:


Fairness Measure
Mid-level Egalitarian Principle(s)
Fundamental Moral Principle(s)
Relevant Empirical Facts

So, to provide an illustrative (if overly simplified) example: supposing the principle of utility is the correct fundamental moral principle, the mid-level egalitarian principle *ensure each person has equal chances at obtaining some good* might be justified in some context in virtue of the fact that satisfying that principle maximizes utility in such contexts; and the use of a fairness measure in such contexts is, in turn, appropriate if and only if it best (or sufficiently) approximates the verdicts of that mid-level egalitarian principle.

There are a few important observations that our general picture helps make clear.

First, it shows that disagreements over a fairness measure can have many sources beyond how a particular measure handles a particular case or how best to decide between equally desirable but mutually unsatisfiable measures. A disagreement over a fairness measure could be the result of a disagreement over what is the correct mid-level egalitarian principle, or it could result from an even deeper disagreement over which fundamental moral principles are the right ones. It could also stem from a disagreement over empirical facts or over how to combine those facts with the fundamental moral principles. Or we could agree on all of this while still disagreeing about which measure best approximates the relevant mid-level egalitarian principle(s).

Second—and relatedly—our picture shows that there are a number of normative criteria to evaluate fairness measures. A choice of a fairness measure could falter in each of the ways that there is a possibility for disagreement. For instance, we might be correct in how to combine the correct fundamental moral principles with the empirical facts to identify the relevant mid-level egalitarian principle(s) for the context, yet we may err in our selection of a measure, misidentifying the appropriate measure for the task (relative to the egalitarian principle that we have identified). We can also, of course, make errors about which fundamental principles are the right ones, about how to combine them with the empirical facts, and so on.

Third, our picture makes clear that defenses and rejections of fairness measures can be extremely limited. Once we realize that a given measure isn't just right or wrong unconditionally, but, rather, is right or wrong given a particular context, we can see that a counterexample or defense of a measure might only apply within a narrow range of cases. In other words, showing that a given measure gets things wrong in a particular case might not show that it isn't a helpful measure in other cases.

Finally, as alluded to above, our picture shows that we do not have to agree all the way down to agree on a measure. People with fundamental moral disagreements might still agree on the appropriateness of the use of a particular fairness measure in some particular context. For instance, it is perfectly possible for a utilitarian and a contractualist to agree that, within a certain context, we should satisfy a particular mid-level egalitarian principle. Even though the contractualist and utilitarian have different reasons for endorsing that mid-level principle, they could converge on a specific egalitarian goal within a certain context and thus both be satisfied by a measure that sufficiently tracks their shared egalitarian goal in that context.[5]

Importantly, we take this structure to be consistent with a wide range of normative theories, and, neither here nor in what follows, will we commit to any particular fundamental or mid-level principles. Instead, we will utilize this structure to illuminate different ways in which various fairness measures might succeed or fail. To this end, we will next explore some of the space of different measures by considering how a few fairness measures fit—or fail to fit—with various mid-level egalitarian principles. The ultimate purpose here is to make good on the claim that there are indeed different measures that are desirable in different contexts, shed some light on the vast space of measures and egalitarian principles we might use, and show how thinking about the egalitarian principles that motivate our choice of measure can help us make at least some progress in selecting a measure appropriate for a particular context.

3. Exploring the Space of Fairness Measures

We will now explore the connections between various mid-level egalitarian principles and context-specific fairness measures. In the case of each pairing, we take it that the measure will—at least in *some* contexts—be perfectly fine to use as part of a regime for ensuring the kind of fairness under discussion. What we are interested in exploring, then, are the robustness of the connections between different egalitarian conceptions of fairness and particular measures of fairness. We will do this by asking two questions about each principle/measure pair: Is satisfying the measure sufficient for satisfying the principle? And is satisfying the measure required for satisfying the principle? We will answer negatively in each case, showing that while a given measure may be useful in some cases, none of the measures we discuss are generally necessary or sufficient for ensuring the sorts

[5] Cf Rawls (1987: 10-11) on the possibility that people with fundamental moral disagreements can reach 'overlapping consensus' on principles to be used in a particular practical context. Cf also Kagan (1997) on the possibility that theories with different fundamental moral commitments can exhibit partial agreement in their first-order principles.

of fairness picked out by the mid-level egalitarian principles under discussion here. This, we take it, is evidence for our claim that no existing fairness measure is context-invariant;[6] rather, there is a plethora of measures to choose from and choosing a measure for a context will take care and great sensitivity to the normative and empirical details of that particular case. We will work from principles and measures that are less demanding to principles and measures that are more demanding, using our counterexamples to motivate the more demanding principles and measures.

Before embarking on that task, though, let us make two further important remarks about what we take success in our task to consist in.

The first remark concerns the fairness measures that we will be considering. These measures are mathematical formulae. Our task would be uninterestingly easy if, in order to show a mismatch between a formula and a mid-level egalitarian principle, we could simply stipulate a case in which someone made an obvious mistake in the use of a formula (e.g., someone who uses a measure that disallows selecting on the basis of some obviously unfair variables but does not take someone's race, say, to be such a variable). Likewise, our task would be uninterestingly difficult if, in order to show a mismatch, we had to demonstrate that, in some given case, it was *impossible* to substitute values for the relevant formula's variables in a way that matched the verdicts of the relevant mid-level egalitarian principle. If the values of the formula's variables could be chosen without restriction, on a case-by-case basis, it would always be possible to find values that would force a 'gerrymandered' match between the formula and any given egalitarian principle. To chart an interesting path between these uninteresting alternatives, we propose to assume that the fairness measure in question is—roughly speaking—being used *as intended*. In other words, we assume a good-faith effort to hook up the measure's variables to one's data set on the basis of a fixed rule, given by the relevant egalitarian principle, to be used across multiple cases. That methodological assumption, we take it, ensures that any mismatches that we find are of *practical* interest. Under this assumption, a mismatch would show that there is a non-superficial problem about using a given fairness measure in the service of promoting a given egalitarian principle—it would show that the measure in question is not a foolproof way of operationalizing the relevant egalitarian principle.

The second remark concerns the egalitarian principles that we will be considering. The literature contains a large number of formulations of each of these egalitarian principles. Thus there are, not only first-order normative disputes concerning which of these principles best reflects a concern for fairness, but also first-order normative disputes intramurally among the adherents of each principle about the most plausible or best-motivated formulation of the principle in question. In section 2, we set aside the ambition to adjudicate first-order normative disputes among these principles. Likewise, our interest here is not in adjudicating this kind of intramural dispute. Instead, we propose to work with a generic or representative characterization of each egalitarian principle and look for mismatches between the principle *so construed* and some particular fairness measure(s). Again, we make this methodological move in order to ensure that our results are of practical interest to users of fairness measures. They show that there is at least a presumptive problem about

[6] Importantly, we will not argue for the *impossibility* of a context-invariant fairness measure. But here let us register our skepticism that there could be one.

drawing justificatory support for the use of a particular fairness measure, in some context, from a particular (family of) egalitarian principle(s).

Finally, we should note that, in keeping with the spirit of our methodology, we have tried to build our cases with sensitivity towards a further complication. The egalitarian principles that we will consider are competing accounts of when and why there is fairness *simpliciter*, in a given situation. But, arguably, some of the fairness *measures* that we consider are designed to test for something narrower than whether or not there is fairness *simpliciter* in a given situation. They are designed, rather, to test for what might be called fairness *in a narrow sense*. A judge, for example, might decide fairly in this narrow sense if she bases her judgment only on the arguments offered by opposing counsel, even if background conditions mean that worse-off defendants have access to worse-quality counsel. Here, there's something unfair *simpliciter* about the situation (in virtue of the background conditions), despite the judge's decision being narrowly fair. There are various ways in which this distinction—between fairness *simpliciter* and fairness in a narrower sense—might be made precise. Given our present purposes, however, we can leave this task aside. First, we've attempted to construct each of our cases in such a way that the fairness or unfairness *simpliciter* is in part explained by some narrow sense in which the relevant decision procedure (which the fairness measure is meant to evaluate) is fair or unfair. This is just to say that we've attempted to construct our cases in such a way that this distinction won't cause trouble. Second, even if, upon reflection, some readers are unconvinced—insisting (for instance) that in some of our cases the situation described is unfair *simpliciter* but the relevant decision procedure's use is nevertheless fair in a narrower sense—our discussion should still be of some interest insofar as it allows such readers to refine their judgments about which mid-level egalitarian principle best reflects the narrow sense of fairness that they take to apply to the use of the relevant decision procedure.[7]

3.1 Formal Equality of Opportunity and Fairness Through Unawareness

We'll begin our exploration of the space by examining the relationship between one mid-level egalitarian principle, formal equality of opportunity, and a fairness measure, fairness through unawareness.

Formal equality of opportunity requires that, "positions and posts that confer superior advantages [...] be open to all applicants. Applications are assessed on their merits, and the applicant deemed most qualified according to appropriate criteria is offered the position" (Arneson, 2015). The ideal of formal equality of opportunity asks that characteristics not relevant to the position or post—such as (at least typically) one's race or sex—do not bar applicants from consideration or figure into decisions.[8] Rather, it asks that applicants are instead assessed on their merits alone.

[7] Although our official position about narrower senses of fairness is, as we mentioned, an agnostic one, in footnote 20 we briefly offer reason to doubt that, if there is fairness in a narrower sense, it is normatively relevant in the cases that we discuss. We thank an anonymous reviewer for pushing us to consider this distinction.

[8] Though, in some contexts, a protected attribute very well could be considered a merit. For instance, Hausman (2014) discusses a case in which race is a merit for a director searching for an actor to play the role of Martin Luther King. But, again, such cases are not typical.

Fairness through unawareness asks that a prediction-based decision system not take as inputs protected attributes—i.e., features of a person that are not to be discriminated against (Grgic-Hlaca et al. 2016; Kusner et al., 2018). So, for example, if a bank is trying to decide to whom to lend by predicting who will make their payments on time, fairness through unawareness asks that the bank not use data about race, sex, or ethnicity in issuing their predictions.

That fairness through unawareness might be attractive from the perspective of formal equality of opportunity—or, at least, that it might initially seem attractive—should be intuitive enough. After all, if we—as regulators—are asked to ensure that banks are satisfying formal equality of opportunity, it would seem reasonable to check that their predictive systems satisfy fairness through unawareness. A bank that is filtering applicants for a loan by race seems to be selecting applicants based on a criterion *not appropriate* to the decision. If a bank fails to meet this low bar, that would seem at the very least to raise a red flag.[9]

Based on this gloss, it seems that fair equality of opportunity and fairness through unawareness at least sometimes make sense as a pair. But now let us ask how robust that connection might be.

Begin with the question of sufficiency: is satisfying fairness through unawareness generally enough for satisfying formal equality of opportunity? We think that it is fairly easy to see that it is not.

To see why, consider the following case:

> **Graduate School.** The admissions system for a graduate program uses an ostensibly race-blind machine learning system to screen applicants; they now receive too many to make the first cut by hand. Unbeknownst to its users and creators, the system was trained on a data set that was heavily biased against a small minority group; further, membership in the group was redundantly encoded in the data set and applications. As a result, certain members of that group have their applications tossed out because the system can identify—and base predictions on—group membership even though the system is never explicitly told who belongs to which group.

In this case, fairness through unawareness might be satisfied, so long as the system uses the test score as an input but does not directly ask about race or ethnicity. Despite satisfying fairness through unawareness, though, such a system would not satisfy formal equality of opportunity because the position is not (in the relevant sense) open to all applicants: certain applicants fare no better than if they had not applied, simply because of their race or ethnicity.

Having seen that satisfying fairness through unawareness isn't in general enough for satisfying formal equality of opportunity, we can turn to the question of whether satisfying fairness through

[9] Cp Arneson (2005): "It should be noted that formal equality of opportunity … puts moral constraints on market decisions ... If one operates a business and provides a product or service to the public for sale, formal equality of opportunity is violated if one refuses to sell to some class of potential customers on grounds that are whimsical … or prejudiced (section 1)."

unawareness is, in general, required for satisfying formal equality of opportunity. To see that it is not, consider the following case:

> **Jobs**. You are hiring. Job applicants take a free aptitude test. You know that members of some minority suffer from a pronounced stereotype threat that reduces their score on this test.[10] (They are just as qualified for the position; the testing environment just has this feature.) So when you assess applications, you take their minority status into account by adjusting their scores.

We can imagine that in this case, you take minority status as an input for deciding who should get an adjustment. In such a case, fairness through unawareness is not satisfied (because the scores take a protected attribute as an input), and yet—so long as the adjustment is appropriately calibrated—formal equality of opportunity very plausibly is satisfied. Thus, fairness through unawareness is not in general required for satisfying formal equality of opportunity.[11]

Importantly, there is nothing especially odd or idiosyncratic about these cases. And thus they show (or at least strongly suggest) that fairness through unawareness and formal equality of opportunity can come apart in a significant range of cases. So if fairness through unawareness is to be an appropriate measure of fairness in some context in which formal equality of opportunity is the salient egalitarian principle, we need some reason to think that the context we are in is not one of those in which they come apart.

3.2 Formal Equality of Opportunity and Equalized Odds

Having seen that fairness through unawareness is in general neither necessary nor sufficient for formal equality of opportunity, let us consider one more fairness measure as it relates to formal equality of opportunity: equalized odds. Equalized odds asks that the rates of true positives and false positives are equal for each group (demarcated by protected attributes). In symbols, equalized odds demands the following, where y stands for, “the subject has the property that the predictor is trying to predict”, $\hat{y}$ for, “the predictor predicts that the subject has the property”, and a for “the subject belongs to such-and-such protected class(es)”:

$$p(\hat{y} \mid y \;\&\; a) = p(\hat{y} \mid y \;\&\; {\sim}a)$$

[10] We should note here that we are not claiming that stereotype threat actually manifests this way. In this case, as in all other hypothetical cases in this paper, we are stipulating plausible—if oversimplified—causal claims strictly for illustrative purposes.

[11] Here, it’s worth emphasizing that the practical force of this conclusion depends in part on whether and to what extent something like the causal dependencies exhibited in our example manifest in the real world. At one extreme, they might represent mere conceptual possibilities and hence lack any practical upshot. This extreme, we reject. As mentioned in footnote 10, we construct our examples to mimic plausible real-world cases in which a fairness measure fails to track salient egalitarian concerns. That said, exactly when and how often examples of the sort we discuss arise in the real world is an empirical question that lies beyond the scope of this paper but whose answer—for precisely the reasons argued for here—is critically important for determining when and why any particular fairness measure should (or should not) be deployed in any particular circumstance.

and

$$p(\hat{y} \mid {\sim}y \;\&\; a) = p(\hat{y} \mid {\sim}y \;\&\; {\sim}a)$$

Formal equality of opportunity and equalized odds might seem to be an intuitive fit because formal equality of opportunity wants us to consider all and consider them on their merits, and equalized odds asks that we not let protected attributes—which (again) are not typically thought of as merits—affect our detection of merits.

Let us now ask our familiar set of questions. Begin with the question of whether satisfying equalized odds is in general sufficient for satisfying formal equality of opportunity. We can see that it is not by considering the following case:

> **Graduate School 2**: The admissions system for a graduate program uses a highly but imperfectly accurate machine learning system to screen applicants. Wanting qualified applicants to have equal chances of success, they tell the system to treat equalized odds as a constraint. In the population it scores, however, members of a disadvantaged minority that values education highly are much more likely to be qualified. To compensate for this while achieving equality of opportunity, the system engages in a kind of leveling down: it *rejects* certain likely-to-be-qualified applicants from that class, so as to keep true and false positive rates in parity across groups.[12]

In such a case, equalized odds is met by design. Yet, formal equality of opportunity is not satisfied because the position is still not open to all—certain members are *rejected* from consideration in virtue of belonging to a protected group.

Having seen that satisfying equalized odds isn't in general enough for satisfying formal equality of opportunity, we can turn to the question of whether equalized odds is in general required for satisfying formal equality of opportunity. We can again show that it is not, this time by considering a variant of Jobs:

> **Jobs 2**. You are hiring. Job applicants take a highly—but not perfectly—accurate, free aptitude test. Members of a disadvantaged minority group are *much* more likely to be qualified. This is because other employers discriminate against this group, leaving a high proportion of qualified members of the group on the job market. As a result of the discrepancies in the base rates—where members of the minority group are much more likely to be qualified—and the test's being highly, but imperfectly, accurate, true positives

[12] This case might be a little confusing. How can we have an artificially high rejection rate for one group while achieving equal true and false positive rates across groups? The key is that (a) the system is not perfectly accurate (creating, in essence, some wiggle room) and (b) the uneven distribution of qualifications which, under certain conditions and in the absence of an artificially high rejection rate for the group or some other corrective (such as perfect accuracy), will reveal itself in the form of asymmetrical true positive or false positive rates. For a more fully fleshed out example of this phenomenon, see section 3.3. For a real-word instance of this phenomenon, see Corbett-Davies et al. (2016)).

> are more common among the minority group than in the majority group (i.e., it violates equalized odds).

Here we have a case where formal equality of opportunity is met and yet equalized odds is not.

Formal equality of opportunity is met in this case because all applicants are being considered and they are being considered on their merits. One might respond by saying that this couldn't be true in the case because the system is—in virtue of the same facts that constitute its violation of equalized odds—biased in favor of detecting qualifications in the minority group. In response, it is enough to note that formal equality of opportunity is compatible with statistical discrimination, so long as the statistical discrimination makes use of the available information in a way that is relevant to appropriate criteria, which it is in this case—the bias is a result of the test's high accuracy and the difference in base rates among applicants.[13]

3.3 Interlude: Formal Equality of Opportunity, Fairness Through Unawareness, and Equalized Odds

So far, we have shown that the verdicts of fairness through unawareness and equalized odds can come apart from those of formal equality of opportunity. But this is consistent with one measure or the other being in general the correct measure for prediction-based decision making (in which case, so much the worse for formal equality of opportunity). And it's also consistent with formal equality of opportunity being in general the correct mid-level egalitarian principle to approximate for prediction-based decision making (in which case, so much the worse for the aforementioned measures). So before we move on to discussing other principles and measures, it's worth pausing to consider whether any we have discussed so far are, as a general matter, the correct principles or measures for prediction-based decision making.

Let's begin by asking—independently of whether they coincide with formal equality of opportunity—if the measures we have discussed seem to get things right in the cases we have discussed so far. It seems that they do not. Each measure approves of a graduate school system, and each system seems unfair to us (in the sense of "unfair" the measure is meant to track). Whatever the correct principle of fairness is (in this context), it seems to be violated in that case. Thus, it isn't true that, as a general matter, satisfying either of these measures is sufficient for being fair.

We can now turn to the question of whether satisfying either measure is necessary for being fair. Again, we think not. Both measures condemn the systems in Jobs and Jobs 2 and, yet, we do not think that the systems in those cases are unfair (in the contextually relevant sense).

Now, this claim might receive some pushback in the case of Jobs 2. Why don't we think that the seeming bias against the majority group is an instance of unfairness? Let's begin by filling in more details of the case. Suppose that the test—while violating equalized odds—satisfies a different fairness measure, *calibration*, which asks that, conditional on our predictions, "outcomes" (our *y*

[13] For further discussion of the compatibility of formal equality of opportunity and statistical discrimination, see Arneson (2015), section 1.3.

variable) are probabilistically independent of protected attributes (Corbett-Davies and Goel 2018). In symbols, calibration asks that

$$p(y \mid \hat{y}\& a) = p(y \mid \hat{y}\& \sim a)$$[14, 15]

and

$$p(y \mid \sim\hat{y}\& a) = p(y \mid \sim\hat{y}\& \sim a)$$

Note that calibration seems to be as intuitive a measure of fairness as equalized odds. But note that calibration and measures like equalized odds are, famously, both seemingly attractive and in tension with one another (see Corbett-Davies and Goel (2018) and Hedden (2021) for helpful overviews of these issues). We can see the tension between the two measures by stipulating some numbers. Suppose the test behaves as follows:

Performance among minority group	Actually qualified	Actually not qualified	Share correctly predicted
Predicted qualified	180	20	90% (of 200)
Predicted not qualified	10	90	90% (of 100)

Performance among majority group	Actually qualified	Actually not qualified	Share correctly predicted
Predicted qualified	45	5	90% (of 50)
Predicted not qualified	10	90	90% (of 100)

The reason these data are helpful is that they show that the test's violation of equalized odds—which we can now quantify: the probability of being identified as qualified if you are a qualified member of the minority group is about 95%[16], whereas the probability of being identified

[14] As before, y stands for, "the subject has the property that the predictor is trying to predict", $\hat{y}$ for, "the predictor predicts that the subject has the property", and a for "the subject belongs to such and such protected class(es)."

[15] We do not have the space here to explore calibration the way we have explored fairness through unawareness and equalized odds, but we think that it—like those scores—is no panacea. To see why, we recommend [REDACTED 1], [REDACTED 2], or Corbett-Davies and Goel (2018).

[16] Since $p(\hat{y} \mid a \,\&\, y) = p(\hat{y} \,\&\, a \,\&\, y) \div p(a \,\&\, y)$ = (number of members of the minority group who are qualified and test qualified ÷ total number of applicants) ÷ (total member of minority group who are qualified ÷ total number of applicants) = $(180 \div 450) \div (190 \div 450) = 180 \div 190 \approx .95$. (Where $a =_{df.}$ "is a member of the minority group").

as qualified if you are a qualified member of the majority group is about 82%[17]—is born out of the fact that the test is both highly but not perfectly accurate and is testing groups with different base rates. The fact that the test gets it right 90% of the time independently of which group one belongs to, in conjunction with the fact that the members of the minority group are much more likely to be qualified (about 65% of the members of the minority group are qualified whereas only about 35% of the members of the majority are), together force this result.

All of this is a long-winded way of saying that the fact that the test in Jobs 2 is more likely to identify qualified minorities as qualified is not a matter of caprice; rather, it is a result of being impartially accurate in a certain kind of way. Now, all of this is not enough to exculpate the test. To draw that conclusion too quickly would be to ignore one of the main lessons of this paper—i.e., that there are no hard and fast rules for measuring machine fairness.[18] Nevertheless, we think that the test in Jobs 2 isn't unfair. Appreciating the fact that the test's accuracy plus the difference in base rates is what causes it to behave as it does helps make clear why.

The further factors that we think make a difference here are as follows. To begin, there is the fact that the test isn't compounding a noxious inequality but in fact ameliorating it. Similarly, the fact that the test ameliorates the inequality in a way that does not impose a cost on the worst off but instead by benefiting them also seems relevant. These facts, in conjunction with the fact that the test is accurate and obeys calibration, justify the deviation from equalized odds. Thus, satisfying equalized odds is not necessary for being, in the relevant sense, fair.

Let us now investigate formal equality of opportunity, which seems to get things right in the cases so far discussed: is it in general the principle of fairness that any measure of fairness should approximate?[19]

Unsurprisingly, we think not. We can see this by considering another graduate school case. Let's start with the question of whether formal equality of opportunity is sufficient for being fair. Considering the following:

> **Graduate School 3**. The admissions system for a graduate program requires scores of a difficult test, and members of disadvantaged minority groups, on average, do poorly on the test because they cannot afford the test preparation needed to be competitive.

[17] Since $p(\hat{y} \mid {\sim}a \,\&\, y) = p(\hat{y} \,\&\, {\sim}a \,\&\, y) \div p({\sim}a \,\&\, y)$ = (number of members of the majority group who are qualified and test qualified ÷ total number of applicants) ÷ (total number of members of the majority group who are qualified ÷ total number of applicants) = (45 ÷ 450) ÷ (55 ÷ 450) = 45 ÷ 55 ≈ .82.

[18] It is here worth noting that there are instances where we think pursuing calibration at the expense of equalized odds is *unfair*. We think that the case that brought to light the tension between calibration and measures like equalized odds—the notorious case of COMPAS (see Angwin et al. 2016 for details)—is one of them. For further exploration of this thought, see [REDACTED 2].

[19] Note that to the extent that the mid-level egalitarian principles are not—like the fairness measures—narrowly focused on a particular *aspect* of fairness, there is a larger stable of potential counterexamples from which we might draw.

In this case, the prediction-based decision system is unfair. Yet formal equality of opportunity does not seem to have been violated: no one is being denied the opportunity to apply in virtue of belonging to a protected group, and the system is—we can stipulate—tracking merits within a reasonable degree of accuracy (relative to the school's one-dimensional goal of admitting the best prepared candidates). Thus, we do not think that formal equality of opportunity is, as a general matter, a correct principle of fairness for prediction-based decision making.

Finally, we can ask whether formal equality of opportunity is required for being fair. To see that it is not, we can consider a modified version of Jobs.

> **Jobs 3**. You are hiring. All are welcome to apply, but minorities are less likely to be qualified due to educational inequities, and members of the privileged majority are more likely to be qualified but not due to hard work or perseverance. As it happens, the job for which you are hiring is one for which some previously unqualified hires were able to learn on the job. So you devise a predictive system for predicting which applicants will *either* arrive qualified *or* be able to learn on the job. You then devise a lottery system to allocate positions among them (so as not to perpetuate inequality).

Of course, the details here matter. But it seems to us that this system could be fleshed out in a way that is fair. Still, formal equality of opportunity is not satisfied because the system does not even attempt to select for the most qualified. Instead, it selects for the sufficiently qualified, in the interest of fairness.[20]

3.4 Substantive Equality of Opportunity and Counterfactual Fairness

In section 3.1 and 3.2 we saw that fairness through unawareness and equalized odds can misfire in cases where the factors that a system treats as qualifications are influenced by subjects' protected attributes. Thus, in all of the graduate school cases, membership of the minority group is causally relevant to whether one is admitted; for some who are denied admission, they would have been admitted were they to have been members of the majority group. In section 3.3, we saw that formal equality of opportunity misfires in similar cases. In Graduate School 3, formal equality of opportunity is satisfied even though the system seems unfair because applicants don't have the same opportunity to develop the skills that the test is testing for.

[20] Arguably, both Graduate School 3 and Jobs 3 raise the complication discussed at the end of section 3: someone might judge that, although these situations are unfair *simpliciter*, the use of these prediction-based decision systems is nevertheless fair in a narrower sense (just as, e.g., judges do not act unfairly in a narrow sense when they base their decisions on the arguments of opposing counsel). Since we are using these cases simply to motivate the interest of the more robust fairness measures and egalitarian principles discussed in sections 3.4 and 3.5, we officially wish to remain agnostic about this issue. But we are doubtful that there is a compelling analogy here with the case of judges. Judges issue their decisions from within a longstanding social practice that there is (we presume) good independent reason to structure in the ways that the legal system is in fact structured. But the design, choice, and use of algorithms in prediction-based decision-making does not, offhand, appear to be embedded in any such social practice. So, even if the fact that some decisions are made fair in a narrow sense by conforming to certain social practices, it is doubtful whether such grounds are available in the case of prediction-based decision-making.

All this suggests that, in certain cases, being fair relies on sensitivity to certain historical facts. The measures and principles we have considered so far are blind to historical facts. Let us, then, consider a principle and a measure that is sensitive to history.

Consider the egalitarian principle *substantive equality of opportunity*, which

> prevails with respect to some desirable position or ranked order of positions just in case all members of society are eligible to apply for the position, applications are fairly judged on their merits and the most meritorious are selected, and *sufficient opportunity to develop the qualifications needed for successful application is available to all* (Arneson, 2015; emphasis ours)

In asking that applicants have equal opportunity to develop qualifications needed for successful application, substantive equality of opportunity brings in a historical element. A formal equal opportunity principle takes the natural distribution of talents or abilities as a given—i.e., beyond the purview of fairness—and is indifferent to whether people can develop these talents. It requires only that, when people's talents or abilities have been developed, careers be 'equally open to (developed) talents'. A substantive equal opportunity principle, on the other hand, though it takes the natural distribution of talents or abilities as a given, is not indifferent to whether people can develop these talents. It thus imposes significant further requirements on what social institutions must do for people, over and above ensuring that careers are open to talents. It is, in this sense, 'substantive'.[21]

Applied to Graduate School 3, substantive equality of opportunity seems to get the right result: applicants have not had the same opportunity to develop the skills necessary for successful application; thus, the test is unfair by the lights of this principle.

Now consider the fairness measure *counterfactual fairness*. Counterfactual fairness says that a prediction-based decision system is fair *if and only if* each of its decisions is the same in the actual world as it is in any counterfactual world where the relevant individual belongs to a different demographic group. To see how counterfactual fairness works, it will be helpful to consider another case:

> **Law School Success**.[22] You are designing a prediction-based decision system for a law school's admissions. What you want to predict is how well applicants will do in their first year. You choose this variable for two reasons. One is that, given the law school's student records, you can predict this with a high degree of accuracy. The other is that first year grades are themselves highly predictive of who will get a job coming out of law school. So, you decide that you will admit only those predicted to get good scores their first year. We

[21] Later, in section 3.5, we consider an even more demanding egalitarian principle that does not take the natural distribution of talents or abilities as a given.

[22] This case and the causal model depicting it are adapted from Kusner et al. (2018).

> can imagine that the system bases its predictions on two inputs: one's undergraduate GPA (grade point average) and LSAT (Law School Admission Test) score.

Now let's ask: is such a system counterfactually fair?

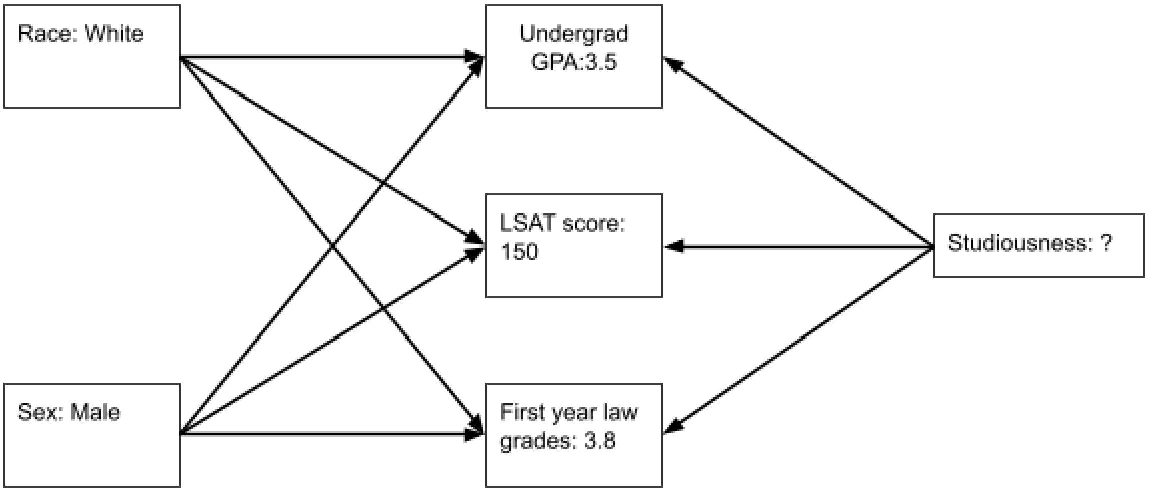
To answer this question, we need to answer a few others. We will first ask what score the system gives a particular individual; we will then need to ask whether they would have received a different score were they to have belonged to a different demographic group.

To do this, let's consider a case. Suppose Jones applies to the school. He applies with an undergraduate GPA of 3.5 and LSAT score of 150. The system finds that applicants with those scores will get a first year average of 3.8, which it considers "good"—the cut-off is 3.7. Now let us turn to the more nebulous question of whether Jones would have gotten a different score were he to have had different protected attributes.

To answer this, we need two things: information about Jones's protected attributes and a causal model—a way of describing how certain features of the world causally interact with each other.[23] We'll stipulate that Jones is a White male and that the causal scenario is summarized by the following model, where items in boxes depict variables and the arrows connecting the boxes depict causal relations (such that an arrow pointing from one variable to another indicates that the former variable is a cause of the latter):[24]

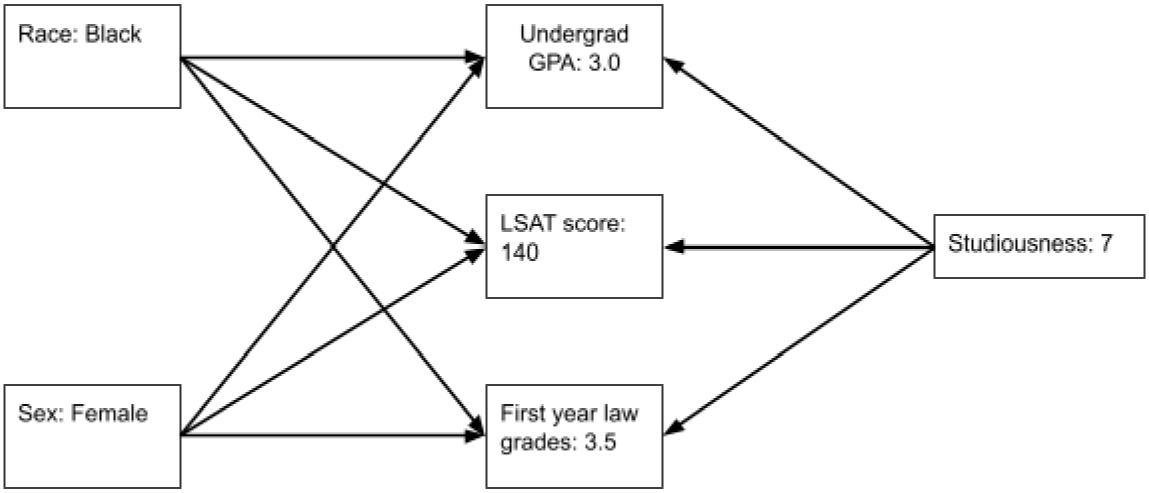
[23] For an accessible primer on causal models, see Pearl et al. (2016).

[24] Note: it's plausible that there are arrows running from race and sex to GPA, LSAT score, and first year law grades since race and sex influence these factors via racism and sexism (via, e.g., differences in class size across schools students are coming from, teachers stereotyping students as more or less likely to succeed or more or less likely to have behavioral problems, and so on). But of course it's not plausible that race and sex are the *only* influences on these things; hence we use "studiousness" as a catch-all to represent the bundle of intrinsic attributes *not* causally related to one's race or sex that influence GPA, LSAT score, and first year law grades. Note also (as previewed in footnote 10) that this is a massively oversimplified model that is being used for illustrative purposes; in reality, there are many more variables than those represented above, and the variables we have chosen might not actually interact exactly in the way pictured here.



With the information about Jones' protected attributes and our causal model, we are in position to ask and answer questions such as "Would Jones have been admitted were he Black and female?". First, we use our causal knowledge and known variables to solve for an unknown variable: studiousness. To keep things simple, we will stipulate the following result: Jones has an average level of studiousness; a 7 on a scale of 10.

Second, we reassign Jones's protected attributes in our model and use our causal knowledge in conjunction with our newfound knowledge of Jones's level of studiousness to determine what Jones' GPA, LSAT score, and (predicted) first year law grades would have been in the counterfactual scenario under consideration. Let us stipulate that the resulting graph summarizes our findings:



Note that in this counterfactual scenario—where Jones is Black and female—Jones is not admitted to the law school (recall that the cutoff for first year law grades is 3.7). For this reason, the system in Law School Success is not counterfactually fair.

This might lead us to wonder what—if anything—is a counterfactually fair way to handle this scenario. The answer is that admitting students on the basis of studiousness—which, in our causal model, is not affected by one's protected attributes—will yield counterfactually fair results. This is precisely because—in the hypothetical world that our model represents—studiousness is not affected by race. Put another way, toggling one's race or sex will never—on this model—toggle their studiousness, so decisions based on studiousness will always pass the test of being counterfactually fair.

Before we turn to our usual set of questions about this measure/principle pair, let us note why one might think that counterfactual fairness and substantive equality of opportunity might make for a good match (in at least some cases). Counterfactual fairness instructs us to base our prediction-based decisions on variables unaffected by membership in protected classes. In many cases, this is a way to turn away from "developed merits" (that applicants may not have had similar chances to develop) and towards variables that applicants will have had similar chances to develop. While—as we will soon demonstrate—the counterfactually fair thing to do won't always be the fair thing to do by the lights of substantive equality of opportunity, there is clearly much to be said for counterfactual fairness's ability to encode substantive equality of opportunity's concerns in particular circumstances.

With this in mind, let us turn to our questions. We'll begin with the question of whether satisfying counterfactual fairness is sufficient for satisfying substantive equality of opportunity, and we will start by scrutinizing the model used in Law School Success.

In our model, neither race nor sex affect studiousness. In the real world, however, it is quite plausible that race or sex affect studiousness. After all, the same racist or sexist factors that affect one's GPA are likely to affect their studiousness. Suppose, for instance, that race indirectly affects studiousness in the following manner. One's race affects their wages, which causes stress that affects their ability to study. So, studiousness is—in the real world—affected by race. So, in a real-world analog of Law School Success, selecting on the basis of studiousness would not be counterfactually fair.

One might think, however, that in any real-life analog of Law School Success it would always be possible to identify a studiousness-*like* variable that is not affected by race—and, therefore, always be possible to use the counterfactual fairness measure in a way that is sufficient to satisfy substantive equality of opportunity. We have two replies. First, even if it is always possible to identify a variable that satisfies counterfactual fairness in this way, there is no guarantee that substantive equality of opportunity would favor selecting on the basis of that variable. Recall that substantive equality of opportunity requires that equally meritorious applicants have similar prospects, irrespective of social circumstances like race or sex. The fact that a counterfactually fair variable is uncorrelated with race or sex does not entail that it is relevant to applicants' meritoriousness. Second, even when it is possible to find a counterfactually fair variable that is, by the lights of substantive equality of opportunity, appropriate to select for, satisfying counterfactual fairness is still not sufficient for satisfying substantive equality of opportunity. To see this, consider another case:

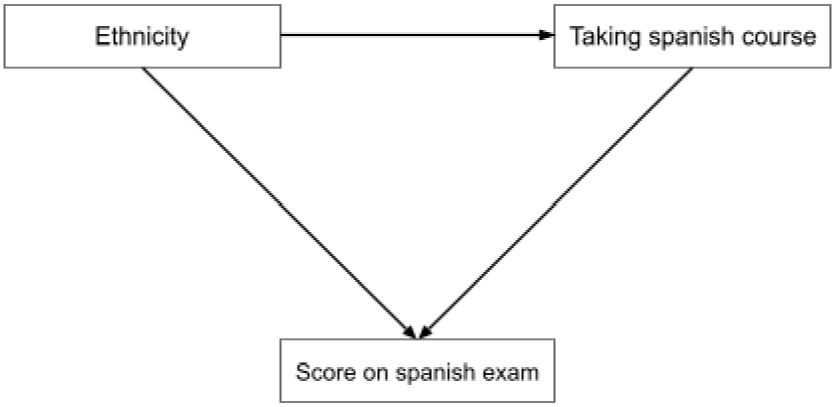
> **Law School Success 2.** Seeing that race affects grades and LSAT score, and valuing counterfactual fairness, the admissions team seeks to base entry on a different variable. As it turns out, applicants come from two different, perfectly integrated local colleges. One happened to begin offering pre-law courses after students in the cohort matriculated, whereas the other didn't. The admissions team uses the pre-law course as a determining factor in their admissions decisions, reasoning that anyone who passed their pre-law courses would excel in law school.

This case shows that there are some deep problems with at least some applications of counterfactual fairness, at least from the perspective of substantive equality of opportunity. After all, if part of the motivation for turning to substantive equality of opportunity was that we thought that it is unfair to judge applicants on attributes they had unequal opportunities to develop, then, in at least some cases—in particular, those in which unequal opportunities don't correlate with demographic group membership—counterfactual fairness is liable to make us unhappy. Suffice to say, satisfying counterfactual fairness is not enough for satisfying substantive equality of opportunity.

Let us, then, turn to the question of whether satisfying counterfactual fairness is required for satisfying substantive equality of opportunity. To see that it is not, consider another case:

> **Internship.** You are hiring for an internship that requires English/Spanish fluency. All students had to take a language class, but only some took Spanish (even though all had the opportunity). Among the cohort, there are many Hispanic students who speak Spanish at home, and thus, are fluent whether or not they took the class. You have applicants take a Spanish exam as part of their application.

Assume that the causal graph for this case looks like this:



Suppose, for simplicity, that taking the Spanish course guarantees that you will score highly enough on the exam to be considered fluent. In such a case, taking the Spanish score as an input runs afoul of counterfactual fairness because one's score on the exam is causally downstream of one's ethnicity—if a Hispanic student who didn't take the course were to have been non-Hispanic, it's possible that that student would not have scored high enough to be considered fluent. Yet, this isn't problematic from the point of view of substantive equality of opportunity, since all applicants had the opportunity to develop Spanish fluency. Thus, satisfying counterfactual fairness is not necessary for satisfying substantive equality of opportunity.

3.5 Luck Egalitarianism and Counterfactual Fairness

Substantive equality of opportunity is a demanding egalitarian ideal. But it is consistent with deep social inequalities that are the result of people having different profiles of 'natural' talents and abilities.[25] Some have therefore proposed a principle that generalizes the considerations that motivate substantive equality of opportunity. This even more demanding *luck egalitarian* principle takes there to be unfairness, not only when people's unchosen *social* backgrounds influence how well their lives go, but when *any* unchosen factors result in lives going unequally well.[26] Luck egalitarianism, then, is the view that it is unfair for some to be worse off than others through no choice of their own.[27]

This preliminary formulation of the view—in particular the 'through no choice of their own' clause—requires further precisification in at least two ways. The first way, well-known to the literature, concerns the interpretation of 'choice'. Someone might become worse off than others in either of two ways. She might be made worse off by some event entirely unrelated to her choices; in the literature's jargon, she might be the victim of bad *brute* luck (which is "a matter of how risks fall out that are not … deliberate gambles"(Dworkin, 2001, p. 73)). Or she might be made worse off by some event that is the unlucky result of an avoidable, risky choice that she made; she might be the victim of bad *option* luck (which is "a matter of how deliberate and calculated gambles turn out—whether someone gains or loses through accepting an isolated risk he or she should have anticipated and might have declined" (Dworkin, 2001, p. 73)). Importantly, 'through no choice of their own' is to be interpreted as condemning only inequalities that reflect bad *brute* luck; luck egalitarianism need not, implausibly, require compensating those who take avoidable gambles and lose out. The second way in which 'through no choice of their own' requires precisification, less well-known to the literature, concerns the interpretation of 'through'. This might be given either a causal reading or a non-causal reading. We return to this matter below, after introducing a case that illustrates the issue.

Unlike formal equality of opportunity (and, in some contexts, substantive equality of opportunity),[28] luck egalitarianism is not a legally codified standard. But luck egalitarianism has been claimed to be

[25] Cf Arneson (2005: section 1); Cohen (2008: 2-3); Segall (2013: vii).

[26] Cf Cohen (2009: 17-18); Hirose (2014: 41-43).

[27] Cf Temkin (1993: 17); Cohen (2008: 7).

[28] See Brighouse et al (n.d.) for discussion, in the case of education, of how substantive equality of opportunity informs the law and other educational standards.

an attractive and plausible principle of fairness in a variety of practical contexts.[29] So, insofar as some users of fairness measures aim to do what, in the relevant context, is fairest, it is worth asking whether any measures of machine fairness reliably track the luck egalitarian view of fairness.

Counterfactual fairness once again seems an obvious fit. We have already seen how counterfactual fairness seems to reflect the considerations that motivate one demanding equal-opportunity principle, substantive equality of opportunity. Thus one might reasonably expect that, by choosing an expanded set of protected attributes—including all those attributes of a person for which it would not be reasonable to hold that person responsible—the counterfactual fairness measure might be used to operationalize luck egalitarianism.[30]

We have already seen, however, reasons to think that satisfying counterfactual fairness is neither necessary nor sufficient for satisfying luck egalitarianism. As noted above, luck egalitarianism is in effect a robust equal-opportunity principle—requiring that how people fare relative to each other be insensitive, not only to unchosen social circumstances, but also to unchosen natural circumstances. And we have already seen, via the Law School Success 2 and Internship cases, that satisfying counterfactual fairness is neither necessary nor sufficient for satisfying such an equal-opportunity principle. Law School Success 2 and Internship already show, then, that satisfying counterfactual fairness is neither necessary nor sufficient for satisfying luck egalitarianism's requirements. However, it is possible to say more, and we think it is worth doing so. As we show in the remainder of this section, given the distinctive motivations for luck egalitarianism, there are some further *kinds* of mismatch between counterfactual fairness and luck egalitarianism, over and above the kinds of mismatch between counterfactual fairness and substantive equality of opportunity discussed above. Since this provides further support for our central thesis—namely, that there is no extant fairness measure that can be used in a context-invariant way to operationalize the concerns expressed by a given mid-level egalitarian principle—we focus in this section on showing that there is a distinctive source or kind of mismatch between counterfactual fairness and luck egalitarianism.

To see, then, a (distinctive) reason to doubt that satisfying counterfactual fairness is sufficient for satisfying luck egalitarianism, consider a somewhat stylized case:

> **Insurance**. A bank offers no-questions-asked loan insurance to businesses. To decide whom to approve, they use an algorithm that is based on a sophisticated, proprietary prediction of the quality of applicants' future business decisions. Frank and Rita have neighboring businesses, both inherited, and neither has any other feasible way of making a living. They

[29] For example, luck egalitarianism has been defended as a plausible approach to ethical questions in healthcare policy (cf Segall (2009) and Eyal (2013)); education policy (cf Voigt (2007) and Harel Ben-Shahar (2016)); and social policy (cf Dworkin (2001) and Segall (2012)).

[30] Indeed, in motivating counterfactual fairness, Kusner et al (n.d.: 7) emphasize this affinity between the measure and the motivations for luck egalitarianism. This apparent motivation for the measure, we take it, is all the more reason to ask how tight is the connection between the measure and the considerations that seem to motivate it. Noteworthy also, as Kusner et al (2018: 3) emphasize, is the counterfactual fairness measure's commitment to facts about individuals, not groups, as the determinant of whether there is unfairness; this is a commitment shared by canonical versions of luck egalitarianism (cf Temkin (1993: 92)).

> both apply for insurance. Frank has a history of foolhardy business decisions, but as it happens (and as the algorithm correctly predicts) will not soon make another one. Rita has a history of responsible business decisions, but as it happens (and as the algorithm correctly predicts) will soon make an uncharacteristically foolhardy one.

The bank thus denies Rita's application and approves Frank's. The algorithm on which they base their decision satisfies counterfactual fairness. It treats people differently only when a factor for which it is reasonable to hold them responsible—the predicted quality of their business decisions—differs. Suppose next that Frank and Rita both default on loans. In Rita's case, this is because of the uncharacteristically foolhardy business decision. But in Frank's case this is because his business is struck by a meteorite, one which could easily have struck Rita's business instead. The bank's actions then leave Rita worse off (her application having been denied) and Frank better off (his application having been approved). But on an orthodox interpretation of luck egalitarianism, we will next argue, this inequality is unfair. And if so, then satisfying counterfactual fairness is not sufficient for satisfying luck egalitarianism.

First, we need to explain why we take it that according to orthodox versions of the luck egalitarian view, the inequality between Frank and Rita is *un*fair. To see the motivation for this claim, consider a hypothetical case that does not involve prediction-based decision making. Two people are badly off, and you can benefit one of them. One of these people is responsible Rick, who is badly off only because of an unpredictable lightning strike. The other badly off person is foolhardy Fionnuala. Fionnuala is badly off as a result of having faced ten identical choices. Each time, she was asked to choose between option 1 (a 99% chance of serious harm and a 1% chance of a moderate benefit) and option 2 (a 99% chance of a trivial benefit and a 1% chance of serious harm). She foolhardily chooses option 1 rather than option 2 the first nine times she faces this choice, and then sensibly chooses option 2 the tenth time. And Fionnuala ends up suffering the serious harm as a result of one of these ten choices.

Now consider two variants of this case: in one, Fionnuala suffers the harm as a result of her *ninth* choice (i.e., the last time she chooses option 1), whereas in two, Fionnuala suffers the harm as a result of her *tenth* choice (i.e., the only time she chooses option 2). Orthodox luck egalitarianism is indifferent between variants one and two of this case. In neither case does it take there to be a reason of fairness to benefit Fionnuala rather than Rick. Relative to the conception of fairness that animates the view, the 'local' facts—i.e., whether some *particular* welfare-reducing event also happens to line up with one's being responsible for that *particular* event's occurrence—are immaterial. To impose this condition would be to impose a "weird causal requirement"[31], one unmotivated by the concerns that animate luck egalitarianism. It is enough, for an inequality to raise luck-egalitarian concern, that it fails to reflect people's 'global' responsibility profiles. This luck-egalitarian orthodoxy seems to us internally well-motivated. Given that Fionnuala makes the very same set of choices in each of the two variant cases, it would be odd to think that it should matter greatly, from the point of view of fairness, whether or not her suffering happens to be causally hooked up to her tenth rather than her ninth choice. After all, *whether* it is the ninth or

[31] Cf Arneson and Hurley (2001: 85)

tenth choice that triggers the harm to her is *not* something within her control. So a version of luck egalitarianism that issued differing verdicts about these two variants of the case would entail that whether an inequality is unfair or not depends on a kind of brute luck. But this sits ill with the standard motivations for luck egalitarianism that we discussed above, which take this kind of luck not to have justifying force.

Now the crucial point that we wish to make here is that the case of Rick and Fionnuala is relevantly similar to the Insurance case. Just as in the case of Rick and Fionnuala, in the Insurance case Rita's being worse off than Frank misaligns with Rita's and Frank's respective 'global' responsibility profiles. And just as in the case of Rick and Fionnuala, it is 'global' responsibility profiles that are, by orthodox luck-egalitarian lights, the morally relevant factor when assessing inequalities. And hence, just as in the case of Rick and Fionnuala, in the Insurance case too Rita's being worse off than Frank would be condemned by orthodox luck egalitarianism.[32]

Having explained the attitude that luck egalitarianism takes toward the inequality in this case, let us now turn to explain why we take it that counterfactual fairness is satisfied in this Insurance case. In making its decision, this firm treats only factors beyond someone's control, and not their responsible choices, as protected attributes; so, by the lights of counterfactual fairness, the firm has not acted unfairly. Now, we have stipulated a version of the case in which Frank and Rita are *first*-time customers of the bank. In other versions of the case, of course, using counterfactual fairness need not come apart from luck egalitarianism. If Frank and Rita had been *long*-time customers, for example, then by appropriately selecting variables, the bank could have ensured that 'global' responsibility considerations drove their decisions about the applications.[33] To see the interest of our result, recall our aim. As we explained in section 3, we are investigating the possibility of mismatches even when the user of a fairness measure 'hooks up the measure's variables to one's data set on the basis of a fixed rule, given by the relevant egalitarian principle, to

[32] To our knowledge, the only dissent from this luck-egalitarian orthodoxy is to be found in Segall (2014: 49-50). (Temkin (2011: 65-66) defends a pluralist view, close enough to the orthodoxy for our purposes, on which 'global' considerations of fairness matter in their own right over and above 'local' ones.) Segall, however, does not intend his discussion to refute the orthodoxy but rather to offer an attractive alternative to it (Segall (2014: 43)). So, for our purposes, this intramural luck-egalitarian dispute can be set aside. The point is that there is a well-motivated luck-egalitarian principle which condemns the inequality in question. The further question of whether, all things considered, this principle is the most plausible version of luck egalitarianism (considered either as an unconditional or mid-level egalitarian principle) is a substantive first-order question of the kind we are setting aside.

[33] There are of course other cases that resemble the Insurance case in which counterfactual fairness would not come apart from luck egalitarianism–for example, cases in which it is a matter of *option* luck whether these business-owners were *in* the situation in which they are exposed to these risks of harm in the first place. We have stipulated away this possibility in the Insurance case, by placing Frank and Rita in situations to which they have no reasonable alternative option but to run their respective inherited businesses, and by stipulating that the inequality is in part the result of a kind of force majeure–namely, a meteorite strike–that could not plausibly be claimed to be an instance of option luck. These details of the case, although somewhat fanciful, illustrate vividly the *kinds* of cases in which this kind of mismatch could arise: one-off interactions between the algorithm and those subject to it, with a great deal at stake, against a background for which no one can reasonably be held responsible, and in which the algorithm tracks 'local' responsibility facts. We think this is a plausible causal possibility, but as we noted in footnote 11, we take no stand on the empirical question of how often such cases in fact arise. (We thank an anonymous reviewer for helpful comments on this point.)

be used across multiple cases'. The Insurance case fits that bill. Indeed, it is hard to see how, in ordinary circumstances, a firm could use the counterfactual fairness measure *other* than by following such a rule. What the Insurance case shows is that mechanically following such a rule—even one that, over time, could be kosher by luck-egalitarian lights—is not always enough. Our point, then—here as in earlier sections—is not that the counterfactual fairness measure is a fundamentally flawed way of operationalizing this mid-level egalitarian principle. Our point, rather, is that users of this measure need to be aware of the possibility of cases that are relevantly similar to our stylized one. If users of the measure are in such circumstances, then the relevant mid-level egalitarian principle would not—as it in other circumstances might—justify using the measure.

Let us now continue on to show that satisfying counterfactual fairness is not necessary for satisfying luck egalitarianism. As noted above, we could here return to the Internship case from §3.4. Recall there, a score on a Spanish-language test was used to decide whom to hire. Now consider the inequality between someone of Hispanic background who is selected for the internship on this basis and someone from a non-Hispanic background who is not. By luck-egalitarian lights, there is no unfairness here, given that the choice not to take the class is one for which the latter person could reasonably be held responsible. Thus, although using the test is not counterfactually fair, it is unobjectionable by luck-egalitarian lights. But, again, there are two further *distinctive* reasons, worth separate mention, to think that satisfying counterfactual fairness is not necessary for satisfying luck egalitarianism. They each illustrate different ways in which, in some cases, to 'mechanically' read off verdicts about fairness from the causal facts would be to overlook considerations to which luck egalitarianism is sensitive.

Consider first another artificially simple case:

> **Drug**. Two doses of a scarce drug are to be distributed among two patients, Alice and Bob, who are not responsible for their health conditions. The hospital makes the decision on the basis of an algorithm that predicts which distribution of doses would deliver the greatest aggregate health benefit. As it happens, that's giving one dose to each, which would restore each to the same intermediate level of health.

Using the algorithm in Drug would not satisfy counterfactual fairness: how much each of these persons stands to benefit from some health intervention is not an attribute for which it is reasonable to hold either responsible. For example, *had* the benefit of giving Bob both doses been greater than the aggregate benefit of giving Alice and Bob each one dose, the algorithm would recommend that Bob get both doses, and clearly this would be unfair by luck-egalitarian lights. But luck egalitarianism is satisfied in this case, for the trivial reason that the algorithm's use, in this context, renders people equally well off. And, according to an influential version of luck egalitarianism, the view only ever condemns (some) welfare *in*equalities but never welfare *e*qualities. (Indeed, luck egalitarians arguably must endorse an 'asymmetrical' view of this kind if they are to preserve the credentials of their view as a genuinely *egalitarian* one—and not one a view that simply reduces to a disguised concern that people should enjoy a level of wellbeing that reflects their responsible choices.[34]) So satisfying counterfactual fairness is not necessary for satisfying luck egalitarianism. What the Drug case shows, in other words, is that the luck egalitarian need not always have an objection to the use of an algorithm that is not counterfactually fair—even if, as in this case, the algorithm favors an end like welfare maximization that generally conflicts with luck egalitarian requirements.

Even if one sets aside such 'asymmetrical' versions of luck egalitarianism, satisfying counterfactual fairness would not be necessary for satisfying luck egalitarianism. For a counterexample of the same kind with a different source, we can return to a variant of our earlier Insurance case:

> **Insurance 2.** The background circumstances are all as in the original Insurance case, except the following. The bank that offers no-questions-asked loan insurance to businesses uses an algorithm that is based on a sophisticated, proprietary prediction of catastrophic events in very localized areas. The prediction engine determines that Frank's business, but not Rita's, is likely to suffer a catastrophic event. And as before whereas foolhardy Frank often makes foolish business decisions, responsible Rita rarely does.

Suppose that, as it happens, again Frank defaults on a loan because his business is struck by a meteorite that could easily have struck Rita's business, and Rita defaults on a loan because of an uncharacteristically foolhardy business decision. The bank's actions would leave Rita better off (since her application would have been approved) and Frank worse off (since his application would have been denied). But, for the reasons explained earlier in this section, orthodox luck egalitarianism would *not* judge this inequality to be unfair. (In this variant of the case, the inequality between Frank and Rita *aligns* with their 'global' profiles of (ir)responsibility.) So, although the algorithm does not satisfy counterfactual fairness (since it would have treated Frank differently, had a factor for which it is not reasonable to hold him responsible—namely, whether his business was afflicted by the catastrophic event—differed), its use here would not be objectionable by luck-egalitarian lights.

The stylized cases discussed in this section are, of course, somewhat unusual. We do not discuss them because we think it especially likely that users of counterfactually fair prediction-based decision procedures would often face such scenarios. Rather, we discuss them because they illustrate in a particularly vivid way a general theme that emerges from our discussion of counterfactual fairness. Counterfactual fairness reads off, from inspection of a prediction-based decision procedure in light of the (assumed) causal facts, the presence of fairness or unfairness. But our discussion shows that both substantive equality of opportunity and luck egalitarianism are, in various ways, sensitive to information *other* than that contained in the (mere) causal facts. That is why counterfactual fairness is not a foolproof guide to fairness, construed luck-egalitarian-wise.

4. Some Lessons

[34] For discussion of these issues and defense of this 'asymmetrical' version of luck egalitarianism (so-called because it treats asymmetrically inequalities and equalities that arise from factors beyond people's control), see Segall (2016: chapter 3). As before, we are not taking a stance on the first-order normative intramural debate among luck egalitarians; the present point is that this 'asymmetrical' version of luck egalitarianism is *a* well-motivated version of the view.

Having completed our tour of an admittedly small fraction of the vast space of fairness measures, let us step back and draw some lessons from what we have done.

One lesson is that choosing proper measures requires nuance and great sensitivity to the egalitarian principles that undergird our choice of measure. This lesson has important applications going forward. Consider counterfactual fairness. Clarifying one's egalitarian motives for pursuing counterfactual fairness might help answer an important question about any pursuit of counterfactual fairness. As the progenitors of counterfactual fairness are keen to note, pursuing counterfactual fairness often comes at a price: loss of accuracy. This loss of accuracy raises a moral concern: how are we to think about trade-offs between fairness and accuracy?[35] While we will not purport to answer this question as a general matter, we do think that our approach offers some guidance. Recall Law School Success 2. Imagine that in using the pre-law course as a variable, we hardly lose any accuracy at all; that is, passing the course turns out to be highly predictive of law school success. Suppose the admissions committee nevertheless asks whether the loss of accuracy in moving from the even more predictive (but counterfactually unfair) studiousness variable is morally acceptable. Our approach gives the admissions committee guidance. We could tell the committee to focus less on how much accuracy they are losing and to focus more on *why* they are losing accuracy and what the proper egalitarian goal has to say about it. If the proper principle for the occasion is substantive equality of opportunity, then the counterfactually fair system that uses the pre-law course variable will not be satisfactory and the committee should look for some other way of designing their system, because substantive equality of opportunity condemns its use. The more general lesson here is that if we bring to bear the egalitarian principles that are driving our choice of fairness measures in the first place, nebulous questions about trade-offs between accuracy and fairness can become better defined questions of whether the systems associated with those trade-offs are acceptable.

The other lesson that we would like to highlight is that—despite our confidence in a kind of pluralism about fairness measures—some general patterns exist, and being aware of these patterns can help in making decisions about which measures to use. Here are four themes from our discussion of principles and scores: 1) As demonstrated by the graduate school cases, fairness through unawareness and equalized odds are liable to miss cases of unfairness where qualifications are caused by protected attributes. 2) As evidenced by Jobs and Jobs 2, fairness through unawareness and equalized odds can misdiagnose cases where protected attributes are good evidence of qualifications. 3) While counterfactual fairness can at least sometimes correct for these shortcomings, it has its own flaws; as Law School Success 2 shows, counterfactual fairness can miss unfairness where qualifications are unfair because, for instance, they are not under subjects' control. 4) As Internship demonstrates, counterfactual fairness can—much like fairness through unawareness and equalized odds—misdiagnose cases where protected attributes are good evidence of qualifications.

[35] For what it's worth, satisfying various other fairness measures also requires (to varying degrees) sacrificing accuracy. And so this question arises, and the following considerations apply, for many other measures in addition to counterfactual fairness.

5. Conclusion

We hope to have shown that there is good reason for those interested in machine fairness to pay more attention to the normative concerns that undergird choices of fairness measures. We also hope to have provided—in the form of our "general picture" and our exploration of the space of measures—some helpful tools and observations for thinking about choices of measures. Finally, we hope to have shown that there are (at least as yet) no magic bullets here—there is not one measure appropriate for all contexts; rather, there are many scores to choose from and much careful work to be done in choosing which one to use in a particular context.

Sources:

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016, May 23). Machine Bias: There's Software Used across the Country to Predict Future Criminals and It's Biased against Blacks. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing

Arneson, R. and Hurley, S. (2001) Luck and equality. *Proceedings of the Aristotelian Society* 75: 51-90.

Arneson, Richard, "Equality of Opportunity", The Stanford Encyclopedia of Philosophy (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2015/entries/equal-opportunity/>.

Binns, Reuben, Fairness in Machine Learning: Lessons from Political Philosophy (December 8, 2017). Conference on Fairness, Accountability, and. Transparency, New York, Forthcoming , Proceedings of Machine Learning Research, Vol. 81, p. 1–11, Forthcoming , Available at SSRN: https://ssrn.com/abstract=3086546

Brighouse, H., Geron, T., and Levinson, M. (n.d.). Conceptions of equity.

Cohen, G. A. (2008). *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.

Cohen, G. A. (2009). *Why Not Socialism?* Princeton, NJ: Princeton University Press.

Corbett-Davies, Sam & Goel, Sharad. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks: It's Actually Not That Clear." Washington Post, October 17, 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

Dworkin, R. (2001) *Sovereign Virtue*. Cambridge, MA: Harvard University Press.

Johnson, G.M. Algorithmic bias: on the implicit biases of social technology. Synthese 198, 9941–9961 (2021). https://doi.org/10.1007/s11229-020-02696-y

Eyal, N. (2013). Leveling down health. In Eyal, N. et al (eds.), *Inequalities in Health: Concepts, Measures, and Ethics*. Oxford: Oxford University Press.

Grgic-Hlaca, Nina, Zafar, Muhammad Bilal, Gummadi, Krishna P, and Weller, Adrian. The case for process fairness in learning: Feature selection for fair decision making. NIPS Symposium on Machine Learning and the Law, 2016

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

Harel Ben-Shahar, T. (2016). Equality in education: why we must go all the way. *Ethical Theory and Moral Practice* 19: 83-100.

Hausman, Dan. “Affirmative Action: Bad Arguments and Some Good Ones.” In Russ Shafer-Landau, ed. *The Ethical Life: Fundamental Readings in Ethics and Moral Problems*, 3rd. ed. New York: Oxford University Press, 2014, pp. 476-489.

Heidari, Hoda, Loi, Michele, Gummadi, Krishna P., and Krause, Andreas (2019). A moral framework for understanding fair ML through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT* '19*). Association for Computing Machinery, New York, NY, USA, 181–190. doi:https://doi.org/10.1145/3287560.3287584

Hedden, Brian (2021). On statistical criteria of algorithmic fairness. Philosophy and Public Affairs 49 (2):209-231.

Kagan, S. (1992) The structure of normative ethics. *Philosophical Perspectives* 6: 223-242.

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

Kusner, Matt J., Joshua R. Loftus, Chris Russell, Ricardo Silva Counterfactual Fairness. Neural Information Processing Systems (NeurIPS), 2018.

Loftus, J., Russell, C., Kusner, M., Silva, R. (n.d.) Causal reasoning for algorithmic fairness.

Pearl, J. et al. “Causal Inference in Statistics: A Primer.” (2016).

Rawls, J. (1987) The idea of an overlapping consensus. *Oxford Journal of Legal Studies* 7(1): 1-25.

Segall, S. (2009) *Health, Luck, and Justice*. Princeton, NJ: Princeton University Press.

Segall, S. (2012) Should the best qualified be appointed? *Journal of Moral Philosophy* 9(1): 31-54.

Segall, S. (2014) *Equality and Opportunity*. Oxford: Oxford University Press.

Segall, S. (2016) *Why Inequality Matters: Luck Egalitarianism, Its Meaning, And Value*. Oxford: Oxford University Press.

Sweeney, L. (2013a) Discrimination in Online Ad Delivery. Available at SSRN: https://ssrn.com/abstract=2208240 or http://dx.doi.org/10.2139/ssrn.2208240

Temkin, L. (1993) *Inequality*. Oxford: Oxford University Press.

Temkin, L. (2011) Justice, equality, fairness, desert, rights, free will, responsibility, and luck. In Carl Knight and Zofia Stemplovska, eds., *Responsibility and Distributive Justice*. Oxford: Oxford University Press. 51-76.

Voigt, K (2007). Individual choice and unequal participation in higher education. *Theory and Research in Education* 5(1): 87-112.