Final version in the *Australasian Journal of Philosophy*

# On the Possibility of Testimonial Justice

Rush T. Stewart  Michael Nielsen

March 26, 2020

**Abstract**

Recent impossibility theorems for fair risk assessment extend to the domain of epistemic justice. We translate the relevant model, demonstrating that the problems of fair risk assessment and just credibility assessment are structurally the same. We motivate the fairness criteria involved in the theorems as appropriate in the setting of testimonial justice. Any account of testimonial justice that implies the fairness/justice criteria must be abandoned on pain of triviality.

**Keywords.** Calibration; credibility; epistemic injustice; equalized odds; fair algorithms

## 1 Introduction

Injustices call for redress. According to a relatively recent literature, some forms of injustice are epistemic. A person offering testimony may fail to be accorded her appropriate level of credibility due to prejudice. More generally, members of salient social groups may tend to have their testimony unduly discounted. Epistemic justice would seem to demand eliminating such biases. But what if eliminating one form of bias makes it impossible to eliminate another? We consider this problem by drawing on a framework from another recent literature on algorithmic fairness.

Consider a widely-discussed case in the context of fair algorithms. The COMPAS risk tool is a statistical method for assigning risk scores in the United States criminal justice system. It is used, for instance, to predict recidivism. Each defendant is assigned a probability of re-offending. An analysis of COMPAS data by ProPublica found systematic bias against black defendants (Angwin et al., 2016). While black defendants were systematically more likely to be incorrectly labeled as higher risk than they actually were, white defendants were more likely to be incorrectly assessed as lower risk than they actually were. The tool's errors, in short, were asymmetric across sub-populations. What are the prospects for doing better?

A growing literature in computer science seeks to address such questions, investigating the possibility of designing fair algorithms for risk assessment (e.g., Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017). Of most interest to this essay, Kleinberg et al. prove that, except in trivial cases, no algorithm can simultaneously satisfy two attractive fairness criteria (2017, Theorem 1.1).[1] One of these criteria is meant to guard against the sort of unfair distribution of errors exhibited in the COMPAS data. If both of their criteria

[1] As an anonymous referee indicated, a similar result is proved earlier in the context of psychometrics and fair selection (Borsboom et al., 2008).

are indeed implied by the appropriate sense of fairness in this context, then risk assessments cannot be fair in general.

In this essay, we are concerned with the prospects for testimonial justice, and just credibility assessments in particular, rather than fair risk scores. Our project is one of model migration. Since mathematical results do not depend on particular interpretations, we can exploit them in other contexts by reinterpreting elements of the model. We will argue that the problem of just credibility assessment can be understood as being formally identical to the problem of fair risk assessment. A plausible translation of the Kleinberg et al. model, we claim, shows that their impossibility theorem applies to testimonial justice: it is impossible, on this interpretation, to avoid testimonial injustice. We show, moreover, that impossibilities remain even for certain ways of weakening the relevant justice constraints. These impossibilities might be interpreted in a number of interesting ways. The most striking conclusion—that testimonial injustice is inevitable in a wide range of situations—is our main focus. But even if one denies this conclusion, we think that our argument still poses some interesting problems for theories of testimonial justice.

## 2 Testimonial Injustice

While it seems little work has been done on formalizing the concept of testimonial justice, there has been a lot of effort to clarify it. In her book on epistemic justice, Miranda Fricker writes, "In face-to-face testimonial exchanges the hearer must make some attribution of credibility regarding the speaker. [...] there can be error in the direction of excess or deficit." (2007, p. 18). Ten pages later, we get a necessary and sufficient condition: "The speaker sustains such a testimonial injustice if and only if she receives a credibility deficit owing to identity prejudice in the hearer" (2007, p. 28).

Fricker makes extensive use of two episodes from fiction to explore the notion of testimonial injustice. In *To Kill a Mockingbird*, Tom Robinson is falsely accused of rape by Mayella Ewell. Tom is black and Mayella is white. The trial comes down to his testimony versus hers. At one point during his testimony, Tom says that he used to help Mayella around the house so much because he felt sorry for her. The prosecutor interrupts, "*You* felt sorry for *her*, you felt *sorry* for her?" In the racist culture in which the novel is set, this is inconceivable for the white prosecutor, judge, and jury *because* Tom is black. As a result, they discount Tom's testimony and the jury convicts him. In *The Talented Mr. Ripley*, Tom Ripley is a conman who murders Dickie Greenleaf. Tom writes a suicide note and makes it look as if Dickie killed himself. Everyone except Dickie's fiancée is convinced that Dickie committed suicide. But Marge knows Dickie well and does not believe that he would do that. She also notices that Tom has Dickie's rings which he promised Marge he would never take off. Marge's resistance to the suicide theory is written off as emotionally motivated. At one point, Dickie's father tells her, "Marge, there's female intuition, and there are facts." According to Fricker, Marge's testimony is dismissed on the basis of the stereotype that women are more emotional and emotionality is at odds with rationality.

There are, naturally, variations in views about how the concepts of testimonial injustice and credibility should be understood. For example, some see credibility *excesses* as more crucial to testimonial injustice than Fricker's definition makes them out to be (e.g., Medina, 2011). According to Jennifer Lackey, attributions of excessive credibility can be harmful in and of themselves. “If I take a black man to be highly knowledgeable about, say, guns or drugs simply because he is a black man, then he has been wronged as a knower just as much as if I take him to be completely ignorant of Shakespeare” (2018, p. 152).

A second important point of variation concerns the extent to which excesses and deficits in credibility must be understood socially. According to Fricker, “credibility is not a good that belongs with the distributive model of justice” (2007, p. 19). On her account, there is enough credibility to go around. So the relevant norm instructs us to attribute as much credibility to an individual as the evidence warrants. Lackey disagrees. Since the account of testimonial injustice that we explore below is of a more social or distributive sort, we quote Lackey at some length.

> Even if you appropriately judge me on the basis of the available evidence and believe accordingly, if you illegitimately regard everyone else as better than I am, I am still the victim of an injustice. Indeed, if others receive a credibility excess, then a credibility deficit to me and an appropriate assessment of my credibility might be functionally equivalent. If this ungrounded asymmetrical treatment pertains specifically to our reports, then I am the victim of testimonial injustice in particular. [...] If you regard my colleague as more reliable than I am, then you will listen to him over me when we disagree, offer him rather than me professional opportunities, and so on. (Lackey, 2018, pp. 154)[2]

In the following section, we propose a framework for modeling and reasoning about distributive aspects of testimonial injustice. In Section 4, we introduce and motivate two desiderata of testimonial justice for a credibility assessor. They are plausibly implied by accounts of distributive testimonial justice. We then show that these desiderata cannot be jointly satisfied (Section 5). One conclusion that could be drawn from our argument is that just credibility assessment is, in general, impossible, though we discuss some other possible conclusions in the final section.

## 3 The Set Up

We begin with a simple example of an assessment problem to aid intuition when we introduce the framework below. The example is one of risk assessment in parole decisions since that’s a standard application. After we introduce the framework, we will explain how it can be interpreted so that it is relevant for credibility assessment as well.

**Example 1.** *Consider a population consisting of six individuals, $A, B, C, D, E,$ and $F$. Suppose that all six people are coming up for parole and that parole decisions are based on predictions about how likely it is that an individual will reoffend. The six individuals fall into two racial categories, Group 1, consisting of $A, B,$ and $C$, and Group 2, consisting of the*

[2] Fricker came to agree that distributive aspects figure into certain legitimate notions of epistemic injustice: “I hope it is clear we may think of the concept of epistemic injustice as an inclusive, generic notion, up for further exploration. In particular, it should be thought of as including distributive forms of epistemic injustice, such as unequal access to epistemic goods like information, or education. In this I agree with David Coady (this issue), who rightly affirms that distributive forms of epistemic injustice are, contrary to what I seem to say at the start of the book, *distinctively epistemic* injustices” (2010, p. 175).

*other three people. In addition to information on race, suppose we also have data on the type of crime for which each person was convicted, with each person falling into one of three categories. Individuals $A$ and $D$ were convicted of the same type of crime, as were $B$ and $E$ and $C$ and $F$. While in reality parole decisions would be made on significantly more information—e.g., employment status before arrest, neighborhood of residence—we will not introduce anything further for simplicity. This population is depicted in Table 1. An asterisk next to an individual indicates recidivism, which is unknown to the assessor at the time of assessment. The task of the assessor $h$ is to predict the probability that an individual is a recidivist. Suppose that for all individuals in Group 1, $h$ predicts a probability of* $2/3$*, and for all individuals in Group 2, $h$ predicts a probability of* $1/3$*. Notice that this assessment has the following nice property. In Group 1, among those with assessment* $2/3$*, the proportion of reoffenders is* $2/3$*. And in Group 2, among those with assessment* $1/3$*, the proportion of reoffenders is* $1/3$*. So, assessments align with actual rates of recidivism within each group—they are well calibrated in a way that we will soon make more precise. (Predicting the group base rate for all individuals in a group is one very simple way to achieve calibration, but there are others.) On the other hand, notice that the assessor also has the following undesirable property. In Group 1, the assessor's expected number of false positives is* $2/3$*, since $B$ is the only individual in that group who will not reoffend, and he receives an assessment of* $2/3$*. But in Group 2, the assessor's expected number of false positives is* $1/3$*, since that is the average assessment of the individuals in that group who will not reoffend. So, the assessor is more error-prone for Group 1 which, in this case, has negative consequences for certain people in that group. Individual $B$, in particular, might complain about his score compared to, say, $E$'s.*

Table 1: Example Population

| | Crime Type 1 | Crime Type 2 | Crime Type 3 |
|---|---|---|---|
| Group 1 | $h(A^*) = 2/3$ | $h(B) = 2/3$ | $h(C^*) = 2/3$ |
| Group 2 | $h(D) = 1/3$ | $h(E) = 1/3$ | $h(F^*) = 1/3$ |

It turns out that the simple assessment task described in Example 1 illustrates a general phenomenon: assessors that are calibrated in the sense suggested by the example are generally more error-prone for one group of individuals than another. Moreover, this phenomenon does not depend on what kind of assessment problem we are considering. It arises for credibility assessment as well as recidivism assessment. With the example in mind, we will now begin to develop this point in a general and precise framework.

Let $N = \{1, ..., n\}$ be a finite set of individuals, the relevant population. In the example, the population consists of six people. Let $Y$ be a random variable taking values in $\{0, 1\}$. Intuitively, $Y(i)$ represents whether or not individual $i$ has a certain property, $y$, with $Y(i) = 1$ indicating that $i$ has $y$, and $Y(i) = 0$ indicating that $i$ does not have $y$. In Example 1, $Y$ is the random variable corresponding to recidivism and represented with an asterisk in Table 1. Let $h$ be a function from $N$ into $[0, 1]$, which we call an *assessor*. The value $h(i)$ represents an assessment about how likely it is that $i$ has property $y$. For instance, if $h(i) = 1$, then individual $i$ is assessed as having property $y$ with certainty. In Example 1, $h(i) = 1$ would represent full confidence that $i$ will reoffend. We will have more to say about how to interpret $h$ in a moment.

Throughout the paper, we will assume that there are just two groups $G_1, G_2 \subseteq N$. For thinking about issues of testimonial injustice, it is natural to consider groupings that carve a population into relevant social identities, e.g., races or genders. Let $P$ be the uniform probability distribution on $N$. Let $P_k = P(\cdot \mid G_k)$ for $k = 1, 2$, so that $P_k$ is the uniform distribution on $G_k$. The quantity $P_1(Y = 1)$, for example, is the proportion of people in $G_1$ with property $y$—2/3 in Example 1. And $P_2(h = 0.8)$ is the proportion of $G_2$ individuals whose assessment by $h$ is 0.8.

Call $\mu_k = P_k(Y = 1)$, $k = 1, 2$, the *base rate* for $G_k$. Let us rule out trivial cases by assuming throughout that $\mu_k \in (0, 1)$. If $k = 1, 2$ and $X$ is a random variable on $N$, then $E_k(X)$ is the expected value of $X$ with respect to $P_k$. For example, $E_1(h)$ is the expected, or average, assessment of $h$ for individuals in $G_1$.

Two aspects of the framework deserve special comment. First, what is the appropriate interpretation of $Y$ if our concern is with credibility and testimonial justice? One possibility is to interpret $Y(i) = 1$ as $i$'s having the property of being *credible* (on a given topic). This interpretation makes credibility a binary variable: each individual is either credible or not. We need not assume that credibility is like this, however. For example, credibility could be interpreted as a graded property that can be coarsened into a binary variable, $Y(i)$, indicating whether $i$ is over some credibility threshold or not. On this interpretation, credibility itself is not binary, but being over or under a particular credibility threshold is. Another possibility is to interpret $Y(i)$ as indicating whether $i$'s report is *true* or *accurate* or *supported by the evidence*. In the rest of the essay, we will use the first interpretation of $Y(i)$, according to which it represents whether or not an individual is credible. What we hope to have indicated here is that not much turns on this. There are a number of other interpretations to which our argument applies.

Second, how are we to interpret the credibility assessments given by $h$? It depends, in part, on how we interpret the variable $Y$, and, as we have just shown, several interpretations of $Y$ are available. One natural interpretation of $h(i)$ is as an assessor's (subjective) probability that $i$ is credible. We use this interpretation in some of the examples because it is an especially intuitive way of thinking about $h(i)$. But it's important to note that nothing hinges on the interpretation of $h(i)$ as a *probability*. Besides taking values between 0 and 1, $h$ is not assumed to have any probabilistic structure (there's no sense in which it is additive, for example). Any interpretation of credibility according to which it can be meaningfully assessed on a ratio scale (normalizable to the unit interval) would do just as well as the probability interpretation: $h(i)$ will just be a credibility assessment of $i$ on that scale. Also, in real life, assessments of credibility about a particular topic will be related to the assessor's overall opinion concerning that topic and to opinions about other topics. In this paper, we remain neutral on how to account for these relations.

## 4 Two Desiderata for Credibility Assessments

The first desideratum for a credibility assessor is calibration. Say that an assessor is *calibrated* if $P_k(Y = 1 \mid h = p) = p$ for all $p \in [0, 1]$ and $k = 1, 2$ such that $P_k(h = p) > 0$. The canonical expository setting for the concept of calibration is weather forecasting. A weather forecaster is said to be calibrated if, for example, it rains on $p$ proportion of the days on which she predicts rain with probability $p$. The forecaster's probabilistic predictions must match the relative frequencies. In other words, it rains on 50% of the days the forecaster predicts rain with 0.5 probability, 70% of the days the forecaster predicts rain with probability 0.7, etc. However, if it rains on 40% of the days the forecaster predicts rain with probability 0.2, the forecaster is said to be *underconfident*. The forecaster is called *overconfident* if, for instance, it rains only on 20% of the days she predicts rain with probability 0.4.

In our setting, calibration requires that among those individuals with a given credibility assessment, the proportion of those who are credible is the same for both groups. It seems fair to us to think that calibration captures one aspect of testimonial justice in this setting. Imposing calibration prevents credibility assessments from being overconfident (or underconfident) in any group. Failures of calibration lead to certain deficits or excesses in credibility across a population, as the following example illustrates.

**Example 2.** *Consider a lab and its sub-populations of female and male scientists. Suppose that the PI's credibility assessments (on a particular topic) of her team are ill-calibrated. Of the males she assesses to be credible at a level of* 0.8*, only half are credible. And of the females she assesses to be credible at a level of* 0.8*, all of them are credible.*

In Lackey's words, "Distributive testimonial injustice, then, occurs, when credibility is improperly distributed among members of a conversational context or community due to prejudice" (2018, p. 157). Example 2 looks an awfully lot like an instance of improperly distributed credibility, with the PI's assessments exhibiting bias against female lab members' testimony and in favor of the testimony of the male lab members. Returning to the language from the weather forecasting setting, the PI's credibility assessments are overconfident in male scientists and underconfident in female scientists. By imposing calibration as a desideratum, we rule out these sorts of inter-group biases.

We recognize that calibration may strike some readers as a rather strong constraint. In Section 5, we consider a couple of ways to relax it. While even these weaker constraints turn out to be problematic for the possibility of achieving testimonial justice, we start with calibration both because it is a relatively familiar property in the philosophical literature on probability and because it is considered "the dominant fairness criterion" in the literature on fair algorithms (Corbett-Davies et al., 2017, p. 799).

The second desideratum concerns a different kind of bias that a credibility assessor might have. To introduce it, we begin with a few definitions. The *false positive rate* of an assessor $h$ for group $G_k$ is $f_k^+(h) = E_k(h \mid Y = 0)$, and the *false negative rate* is $f_k^-(h) = E_k(1-h \mid Y = 1)$.[3] To explain these quantities intuitively, think of $1-h$ as an assessor of *discredibility*. Then, the false positive rate $f_k^+(h)$ is the average credibility assessment of $h$ among individuals in $G_k$ who are *not* credible, and the false negative rate $f_k^-(h)$ is the average discredibility assessment among individuals in $G_k$ who *are* credible. Say that an assessor $h$ exhibits *equalized odds* if $f_1^+(h) = f_2^+(h)$ and $f_1^-(h) = f_2^-(h)$. Equalized odds is the second desideratum for a credibility assessor. It requires error rates to be the same across groups so that neither type of error disproportionately affects one group. For instance, under equalized odds, the average

[3] Recall that the conditional expectation $E_k(h \mid Y = 0)$ is the expected value of $h$ with respect to the conditional probability $P_k(\cdot \mid Y = 0)$ (similarly, for $E_k(1 - h \mid Y = 1)$).

credibility assessments for non-credible individuals are the same for individuals in all groups. Put another way, equalized odds requires that individuals of the same credibility, regardless of group, are treated the same by the assessor in the sense of having the same expected credibility assessment. To take another example of a sort of formal credibility assessment setting, in aptitude testing, equalized odds would require that individuals with the same aptitude have the same expected score on the test regardless of their group membership. If individuals of one race, for example, had a lower expected score than individuals of another race but of the same actual aptitude, the test could rightly be called biased.

Equalized odds fails in Example 1. Here is another example that suggests that failures of equalized odds give rise to testimonial injustice.

**Example 3.** *Consider the sub-populations of black and white witnesses coming before a Maycomb, Alabama jury. Suppose that the jury is calibrated: $p$ proportion of those witnesses in each group that the jury assesses to be credible with probability $p$ are credible. But suppose that the jury's assessments fail to exhibit equalized odds. Non-credible white witnesses, say, receive higher credibility assessments on average than non-credible black witnesses. Or credible black witnesses receive lower credibility assessments on average than credible white witnesses do.*

As Lackey points out, if credible black witnesses receive lower credibility assessments on average than their white counterparts, even if the black witnesses receive the credibility warranted by the evidence, this may be functionally equivalent to the black witnesses suffering testimonial injustice via a credibility deficit.

## 5 Impossibility Results

We will say that a credibility assessment problem is *trivial* if either the base rates for the two groups are the same, $\mu_1 = \mu_2$, or the problem allows for error-free assessment for both groups, $f_k^+(h) = 0 = f_k^-(h)$ for $k = 1, 2$. It seems clear to us that we should expect neither of these conditions to hold very generally. Only in very special cases would the base rates be the same for two interestingly different sub-groups of a general population.[4] Even in the setting of a lab or academic department, credibility will naturally vary on particular topics, and there is no easy, general guarantee of equal base rates for any social groupings. Similarly, in assessment and prediction problems of any significant complexity, perfect prediction is hardly realistically attainable, as the COMPAS data vividly illustrates. If both error rates are 0 for an assessor, then the assessor assigns 1 to all individuals having the relevant property and 0 to all individuals lacking it. But even if most credibility assessment problems of interest to researchers working on testimonial injustice turned out to be trivial in our sense, it would still be unsatisfactory, from a theoretical point of view, to ignore non-trivial cases entirely. An adequate theory of testimonial injustice should not be based on the assumption that groups

[4] Concerning the base rates for credibility in particular, some standpoint theorists argue that "those who are subject to structures of domination that systematically marginalize and oppress them may, in fact, be epistemically privileged in some crucial respects" (Wylie, 2013, p. 26). And, generally, there may be good reason to think population base rates differ with respect to their credibility on certain topics. Consider the population of Alabama partitioned into the set of climate scientists and its complement. There can be little doubt that credibility base rates differ for these groups on topics in climate science.

always have the same proportions of credible individuals, nor the assumption that error-free credibility assessment is always possible.

But now, allowing for non-trivial credibility assessment problems, a serious conflict arises between the two desiderata that we introduced above. We present this conflict as an impossibility result.

**Theorem 1.** *There is no non-trivial credibility assessment problem that allows for a calibrated credibility assessor that exhibits equalized odds.*[5]

In other words, in order to achieve both calibration and equalized odds, the population must consist of two groups whose proportions of credible individuals are exactly the same, or the assessor must be able to determine whether individuals are credible without error. Since, we contend, most realistic and interesting credibility assessment problems will involve groups with different base rates and imperfect methods of assessment, the theorem shows that failures of calibration and equalized odds are to be expected. Insofar as calibration and equalized odds are conditions required for testimonial justice, the theorem shows that testimonial justice cannot be fully attained in non-trivial scenarios.

We now consider three objections. First, one might think that just credibility assessments require both calibration and equalized odds but also think that these are not all there is to testimonial justice. The theorem still applies to views like this because it shows that any account of testimonial justice that *implies* both calibration and equalized odds—sees them as features that should be satisfied by assessments of credibility, even if those assessments should have other properties as well in order to qualify as just—stands in need of revision, since these conditions cannot be simultaneously satisfied in non-trivial cases.

Second, in the literature on fair algorithms for risk scores, calibration, as has been mentioned, is considered "the dominant fairness criterion" (Corbett-Davies et al., 2017, p. 799). But one might deny that testimonial justice by itself requires calibration of an assessor. We now investigate two ways of weakening calibration and show that relevant impossibility results can still be derived. Calibration implies a particular functional relationship between base rates and error rates, as Lemma 1 in the Appendix attests. We might, more simply, assume a systematic relationship between base and error rates directly. For example, let $F : [0, 1] \to \mathbb{R}$ be an injective function such that for $k = 1, 2$ we have

$$F(\mu_k) = \begin{cases} \frac{f_k^-(h)}{f_k^+(h)} \text{ if } f_k^+(h) > 0; \\ \\ \frac{f_k^+(h)}{f_k^-(h)} \text{ otherwise.} \end{cases} \tag{$*$}$$

The piecewise definition of $F$ is required only to avoid 0 in the denominator of the error ratio. Property ($*$) is well-defined assuming the part of non-triviality that excludes perfect prediction. What ($*$) says is that differences in error rates have to be (uniquely) determined by base rates. If they are not, then we could have two populations with the same base rates of some property (e.g., being a recidivist or being credible) yet the assessor identifies more false positives in one population than in the other. Excusing the bias due to a higher frequency of the property in the relevant population is explicitly ruled out. The assumption of injectivity

[5] For completeness, we provide an elementary proof of Theorem 1 in the Appendix.

means that a given error ratio is licensed by only one base rate. We note that, under the non-triviality assumptions, calibration implies the existence of such an $F$. Property ($*$) allows us to state the following generalization of Theorem 1.

**Theorem 2.** *There is no non-trivial credibility assessment problem that allows for a credibility assessor satisfying ($*$) that exhibits equalized odds.*

As is clear from the proof, property ($*$) implies that even simultaneously weakening equalized odds to equal error *ratios*—$f_1^-(h)/f_1^+(h) = f_2^-(h)/f_2^+(h)$—opens up no new possibilities.

Consider now a second way of relaxing calibration. Certain failures of calibration might not imply testimonial injustice, provided the failures are the same for all groups. For example, if the PI's assessments in Example 2 were such that all of the men *and* all of the women she assessed as credible at a level of 0.8 were credible, then her assessments, though ill-calibrated, display the same underconfidence in both groups. The inaccuracy of her judgments affects neither group disproportionately. So there seem to be important differences between different types of calibration failures as far as justice is concerned. These differences derive from the fact that calibration is equivalent to the conjunction of two principles, one relating to equity, the other to accuracy. The first we call *predictive equity*:

$$P_1(Y = 1|h = p) = P_2(Y = 1|h = p)$$

for all $p$ for which the conditional probabilities are well-defined. Predictive equity can plausibly be motivated on purely *justice*-based considerations. The condition requires merely that groups be treated the same in assessment and is consistent with failures of calibration. A purely justice-based motivation is arguably less plausible for the further, accuracy-based constraint that $P_k(Y = 1|h = p)$ equal a particular value, namely, $p$. Furthermore, it is easy to find simple examples of non-trivial assessment problems in which both predictive equity and equalized odds are satisfied. So one way of skirting the impossibility result in Theorem 1 is to relax calibration to predictive equity.

There are two limitations to this approach, however. First, for an interesting class of assessors, this strategy of avoiding the impossibility result will not work. Say that $h$ is a *binary* assessor if it only assesses individuals as credible ($h(i) = 1$) or not ($h(i) = 0$). Binary assessments of credibility are the sort of quick and dirty assessments that we might make in what Fricker calls "face-to-face testimonial exchanges." For binary assessors, predictive equity and equalized odds are incompatible except in trivial cases and for maximally error-prone assessors. We call an assessor *maximally error-prone* if $f_k^+(h) = 1$ and $f_k^-(h) = 1$ for $k = 1, 2$.

**Theorem 3.** *There is no non-trivial credibility assessment problem that allows for a binary assessor that satisfies predictive equity and equalized odds unless the assessor is maximally error-prone.*

This is a simple corollary of a result due to Chouldechova (2017, p. 157), but we provide a proof in our setting in the Appendix.

The second limitation—which applies equally to property ($*$)—is that accuracy may be a desirable feature for credibility assessors even if it is not required by testimonial justice. Calibration is still a plausible and widely-endorsed *epistemic* requirement (e.g., Dawid, 1982;

van Fraassen, 1983; Shimony, 1988; Lange, 1999; Tetlock, 2005).[6] On this way of viewing things, Theorem 1 shows that there is a conflict between requirements arising from considerations of justice and those arising from epistemic considerations: an assessor cannot both exhibit the sort of justice formalized by equalized odds *and* be calibrated. So, Theorem 1 would still seem to be bad news for the prospects of testimonial justice because it would show that justice is not achievable without epistemic sacrifice. Given that such trade-offs are inevitable, it would be nice to have some guidance on how they should be made. One idea from the psychometrics literature is to distinguish tests (assessments) as contests from tests as measurement (Borsboom et al., 2008, pp. 86–87). Aptitude tests, for example, can be used to decide who gets to attend Harvard (contest), but they can also be used to discern the best course of education for a particular person (measurement). In the case of measurement, using all prior information including group membership data may be unproblematic or even required. But it could be quite ethically problematic to use group membership data in the context of a contest. For instance, two students could have the same test score, but the base rate for completing the program is lower for one relevant group than the other.[7] Borsboom et al. quote Howard Wainer: "Measurements must be as accurate as possible. Contests must be as fair as possible" (2008, p. 87). The central idea here is that, whereas a premium is placed on accuracy in the context of measurement, in contest settings, epistemic sacrifice may very well be warranted.[8]

A third objection is that failures of calibration or equalized odds do not imply testimonial injustice when such failures are not due to an identity prejudice conceived of as an intentional state. For example, the PI's assessments are unjust only insofar as they are driven by (explicit or implicit) sexism and not something else. There are at least two categories of reply. First, and somewhat controversially, some think that biased intentional states *can* be read off from disparate group treatment, at least under certain circumstances. Consider claims about the morally problematic nature of the gender wage gap or racial income inequality. On such views, the probability of explicit bias given distributive facts can be rather high. However, a general point that is widely-appreciated from Schelling's (2006) work is that inferring micro-motives from observed outcomes can be a tricky affair.

Second, and more promisingly, it is possible to conceive of distributions as unfair or unjust without tracing the injustice back to the biased intentional states of an individual. Consider, for example, claims about the morally problematic nature of general income inequality in the United States. We need not take a stance on the first type of reply involving inferring biased

[6] But see (Seidenfeld, 1985), for instance, for some critical remarks concerning calibration as a general norm for subjective probability.

[7] At the extreme end of reliance on group base rates is profiling.

[8] There are nuances here that we cannot fully address in this paper. Epistemic sacrifice in contest settings can itself lead to harmful effects on already disadvantaged groups. For example, as Borsboom et al. write,

> [C]onsider again the scenario as it may occur in selection situations that involve minority groups. As explained [. . . ], there are empirical reasons to consider the possibility that accepted members from some minority groups may include a larger number of false positives, which may be partly responsible for the observed increased dropout rates among members of such groups. The presence of more false positives among minorities may create a perceived empirical basis for prejudice. (2008, p. 87)

The precise distribution of the benefits and harms of testimonial injustice, it seems to us, is largely context-dependent. Our primary goal, however, is to indicate substantive constraints for theories of testimonial justice rather than to analyze this context-dependence.

motives. We want to claim only that the model is capable of capturing a morally significant kind of identity prejudice in terms of an assessor's distribution of credibility across the relevant social identities. That is, for present purposes we take the relevant sort of identity prejudice to be a fact about the distribution of credibility for the identity groups involved. If an assessor fails to satisfy equalized odds, for instance, then it is more error-prone for one group than for the other. That, we claim, is one legitimate sense of unfairness or injustice.

Having offered replies to these objections, we think that the results above place interesting limitations on theorizing about testimonial injustice. While we recognize that identifying paths towards potential solutions to these issues would be a valuable contribution, our effort here is directed at identifying the relevant problems in the first place.

## 6 Conclusion

A pessimistic reaction to the impossibility results focuses on the fact that epistemic injustice is unavoidable except in trivial cases. On August 21, 2018, *The Washington Post* ran a story with the headline "Facebook is rating the trustworthiness of its users on a scale from zero to 1" (Dwoskin, 2018). The assessment is part of Facebook's efforts to combat "fake news." Provided the assessment problem is not a trivial one—and it is not—testimonial injustice as it has been construed in this essay is a forgone conclusion.

A less pessimistic reaction construes the upshot of the foregoing limitative results in analogy with Arrow's Impossibility Theorem for social choice (Arrow, 1951). Arrow's theorem establishes that four properties with a great deal of intuitive appeal—properties like the absence of a dictator of social preference and the social adoption of unanimously held individual preferences—are impossible to jointly satisfy. This result has served to structure further theorizing about social and democratic choice, giving rise to careful analyses of related possibilities and impossibilities and normative arguments for and against candidate properties of social choice functions.

More optimistically still, one might deny that the limitations expressed by Theorem 1 apply to testimonial justice properly construed. That is, testimonial justice does *not*, in fact, require calibration and equalized odds. Even if this objection can ultimately be sustained, we hope to have indicated that theorizing about epistemic justice stands to gain from interaction with the burgeoning literature on fair algorithms.[9] From this literature, we have imported a formal model that may serve as a precise setting in which to investigate issues of testimonial justice in a general and rigorous way. While we have attempted to motivate the desiderata that figure in the impossibility results as appropriate for just assessments of epistemic credibility, we recognize that there may well be other accounts. If this is the case, we would like to issue something of a constructive challenge. Exactly what properties should a just credibility assessor have?[10]

[9] This is the literature that triggered our thinking about the subject of this paper, but other established literatures are also relevant, e.g., work in psychometrics on fair selection (Borsboom et al., 2008), and work in labor economics on statistical discrimination (Fang and Moro, 2011).

[10] We were first introduced to the literature on fair algorithms and the sorts of limitative results discussed here in Tina Eliassi-Rad's talk at the Decision & AI conference in Munich in July, 2018. We learned about Jennifer Lackey's views on testimonial justice from episode 98 of the *Elucidations* podcast shortly after. Thanks to Lackey for sharing a draft of her paper with us. And thanks to Jean Baccelli, Patrick Klösel, Alex Reutlinger, Jan-Willem Romeijn, Joe Roussos, Georg Schollmeyer, Shanna Slank, Tom Sterkenberg, Reuben

# Appendix

## Proof of Theorem 1

Following Pleiss et al. (2017), which rehearses the Kleinberg et al. result, the key to proving Theorem 1 is the following lemma, which shows that the error rates of calibrated assessors are linearly related by a coefficient determined by the base rate.

**Lemma 1.** *Let $h$ be a calibrated predictor. If $k = 1, 2$, then*

$$f_k^-(h) = \frac{1-\mu_k}{\mu_k} f_k^+(h). \tag{1}$$

*Proof.* Let $p_1, ..., p_m$ be an enumeration of $h$'s values. Let $S_k \subseteq \{p_1, ..., p_m\}$ be the support of $P_k \circ h^{-1}$, $k = 1, 2$. We start by using calibration to observe that

$$\mu_k = P_k(Y=1) = \sum_{p \in S_k} P_k(Y=1 \mid h=p) P_k(h=p) = \sum_{p \in S_k} p P_k(h=p) = E_k(h). \tag{2}$$

Next, we use Bayes' theorem and calibration to compute

$$\begin{aligned} P_k(h=p \mid Y=1) &= \frac{P_k(Y=1 \mid h=p)P_k(h=p)}{P_k(Y=1)} \\ &= \frac{pP_k(h=p)}{\mu_k} \end{aligned} \tag{3}$$

and

$$\begin{aligned} P_k(h=p \mid Y=0) &= \frac{P_k(Y=0 \mid h=p)P_k(h=p)}{P_k(Y=0)} \\ &= \frac{(1-p)P_k(h=p)}{1-\mu_k}. \end{aligned} \tag{4}$$

Then, we use (3) and (4) to compute

$$\begin{aligned} f_k^-(h) &= 1 - E_k(h \mid Y=1) \\ &= 1 - \sum_{p \in S_k} pP_k(h=p \mid Y=1) \\ &= 1 - (\mu_k)^{-1} \sum_{p \in S_k} p^2 P_k(h=p) \\ &= 1 - (\mu_k)^{-1} E_k(h^2) \end{aligned} \tag{5}$$

Stern, Greg Wheeler, audiences at FEW 2019 and the 2019 Summer School in Mathematical Philosophy for Female Students, and two anonymous referees for very constructive feedback and suggestions that helped to improve the paper a great deal.

and

$$\begin{aligned} f_k^+(h) &= E_k(h \mid Y = 0) \\ &= \sum_{p \in S_k} p P_k(h = p \mid Y = 0) \\ &= (1 - \mu_k)^{-1} \sum_{p \in S_k} p(1 - p) P_k(h = p) \\ &= (1 - \mu_k)^{-1}(E_k(h) - E_k(h^2)) \end{aligned} \tag{6}$$

Now, multiply (6) by $(1 - \mu_k)/\mu_k$ and use (2) and (5) to get

$$\frac{1 - \mu_k}{\mu_k} f_k^+(h) = (\mu_k)^{-1}(E_k(h) - E_k(h^2)) = 1 - (\mu_k)^{-1} E_k(h^2) = f_k^-(h),$$

which is the desired result. □

*Proof of Theorem 1.* Suppose for contradiction that the prediction problem is non-trivial, exhibits equalized odds, and is calibrated. By calibration, Lemma 1 gives

$$f_1^-(h) = \frac{1 - \mu_1}{\mu_1} f_1^+(h) \text{ and } f_2^-(h) = \frac{1 - \mu_2}{\mu_2} f_2^+(h). \tag{7}$$

By non-triviality and equalized odds, there exist $x, z \in (0, 1]$ such that $f_1^-(h) = x = f_2^-(h)$ and $f_1^+(h) = z = f_2^+(h)$. So, by (7),

$$x = \frac{1 - \mu_1}{\mu_1} z \text{ and } x = \frac{1 - \mu_2}{\mu_2} z. \tag{8}$$

Thus,

$$\frac{1 - \mu_1}{\mu_1} = \frac{1 - \mu_2}{\mu_2}. \tag{9}$$

Equation (9) implies that $\mu_1 = \mu_2$. This contradicts our assumption that the assessment problem is non-trivial. □

## Proof of Theorem 2

Interestingly, the weaker property $(*)$ simplifies the proof of Theorem 1.

*Proof.* Suppose for contradiction that the prediction problem is non-trivial and satisfies $(*)$ and equalized odds. By equalized odds, we have $f_1^-(h) = f_2^-(h)$ and $f_1^+(h) = f_2^+(h)$. By non-triviality, we know at least one of $f_k^-$ or $f_k^+$ is positive for $k = 1, 2$. Assume the latter without loss of generality. Using $(*)$, we have

$$F(\mu_1) = \frac{f_1^-(h)}{f_1^+(h)} = \frac{f_2^-(h)}{f_2^+(h)} = F(\mu_2). \tag{10}$$

Since $F$ is injective, (10) implies that $\mu_1 = \mu_2$, contradicting the assumption that the assessment problem is non-trivial. □

Notice that, rather than assuming equalized odds, we could have simply assumed the displayed equality above for error ratios. In other words, simultaneously relaxing both calibration—to (∗)—and equalized odds opens up no new possibilities. Equal error ratios on its own might be unattractive since it allows for the possibility of "compensating" a drop in one type of error in one group with an increase in the other type of error for the other group.

## Proof of Theorem 3

Suppose the assessor is not maximally error-prone and suppose for contradiction that the prediction problem is non-trivial and exhibits predictive equity and equalized odds. To begin, one can verify the following equation, for $k = 1, 2$, by applying the definition of conditional probability and cancelling like terms.

$$(1 - \mu_k)P_k(Y = 1 \mid h = 1)P_k(h = 1 \mid Y = 0) = \mu_k P_k(Y = 0 \mid h = 1)P_k(h = 1 \mid Y = 1) \quad (11)$$

Since $h$ is binary, equalized odds reduces to

$$P_1(h = 1 \mid Y = 0) = P_2(h = 1 \mid Y = 0) \text{ and } P_1(h = 0 \mid Y = 1) = P_2(h = 0 \mid Y = 1), \quad (12)$$

the latter of which implies

$$P_1(h = 1 \mid Y = 1) = P_2(h = 1 \mid Y = 1). \quad (13)$$

By non-triviality the conditional probabilities in (12) are non-zero, and because the assessor is not maximally error-prone, they are not equal to 1. The latter fact implies that the conditional probabilities in (13) are non-zero. Next, predictive equity implies

$$P_1(Y = 0 \mid h = 1) = P_2(Y = 0 \mid h = 1) \text{ and } P_1(Y = 1 \mid h = 1) = P_2(Y = 1 \mid h = 1). \quad (14)$$

The conditional probabilities in (14) are non-zero: if they were zero, then so too would be the conditional probabilities in (12) and (13). Finally, by (12), (13) and (14) we see that the terms in (11) are positive and the same for both groups. Therefore (9) holds, which delivers the contradiction that $\mu_1 = \mu_2$. □

## References

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

Arrow, K. J. (2012, originally published in 1951). *Social Choice and Individual Values*. Martino Publishing.

Borsboom, D., J.-W. Romeijn, and J. M. Wicherts (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods 13*(2), 75–98.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*(2), 153–163.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association 77*(379), 605–610.

Dwoskin, E. (2018, August). Facebook is rating the trustworthiness of its users on a scale from zero to 1. https://www.washingtonpost.com/technology/2018/08/21/facebook-is-rating-trustworthiness-its-users-scale-zero-one/?noredirect=on&utm_term=.aaa0972fc65f.

Fang, H. and A. Moro (2011). Theories of statistical discrimination and affirmative action: A survey. In J. Benhabib, A. Bisin, and M. O. Jackson (Eds.), *Handbook of Social Economics*, Volume 1, pp. 133–200. Elsevier.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford University Press.

Fricker, M. (2010). Replies to Alcoff, Goldberg, and Hookway on epistemic injustice. *Episteme 7*(2), 164–178.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Lackey, J. (2018). Credibility and the distribution of epistemic goods. In K. McCain (Ed.), *Believing in Accordance with the Evidence: New Essays on Evidentialism*, Volume 398, pp. 145–168. Cham: Springer Verlag.

Lange, M. (1999). Calibration and the epistemological role of bayesian conditionalization. *The Journal of Philosophy 96*(6), 294–324.

Medina, J. (2011). The relevance of credibility excess in a proportional view of epistemic injustice: Differential epistemic authority and the social imaginary. *Social Epistemology 25*(1), 15–35.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689.

Schelling, T. C. (2006). *Micromotives and Macrobehavior.* WW Norton & Company.

Seidenfeld, T. (1985). Calibration, coherence, and scoring rules. *Philosophy of Science 52*(2), 274–294.

Shimony, A. (1988). An adamite derivation of the principles of the calculus of probability. In *Probability and Causality*, pp. 79–89. Springer.

Tetlock, P. E. (2017, originally published in 2005). *Expert Political Judgment: How Good Is It? How Can We Know? (New Edition).* Princeton University Press.

van Fraassen, B. C. (1983). Calibration: A frequency justification for personal probability. In R. Cohen and L. Laudan (Eds.), *Physics, Philosophy, and Psychoanalysis*, pp. 295–319. Springer.

Wylie, A. (2013). Why standpoint matters. In *Science and Other Cultures*, pp. 34–56. New York: Routledge.