Equalized Odds is a Requirement of Algorithmic Fairness

David Gray Grant
University of Florida
Jain Family Institute

1 Introduction

In 2016, ProPublica published an analysis of Equivant Inc.'s COMPAS recidivism prediction instrument, arguing that the algorithm was unfairly biased against black defendants. Their analysis found that COMPAS made mistakes at similar rates for black and white defendants. However, it tended to make different *kinds* of mistakes for the two groups: whereas the false positive rate was significantly higher for black defendants than white defendants, the false negative rate was higher for white defendants than black defendants. Since COMPAS was significantly more likely to "falsely flag black defendants as future criminals," ProPublica concluded that COMPAS is "unfairly biased against blacks."[1]

ProPublica's argument implicitly appealed to two putative conditions on the fairness of predictive methods used to evaluate recidivism risk, Equal False-Positive Rates and Equal False-Negative Rates:

> Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.
>
> Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.[2]

The conjunction of these two conditions is called "Equalized Odds."[3] Equalized Odds and its component criteria are examples of *statistical criteria of fairness.* Statistical criteria of fairness are conditions on the fairness of predictive methods that can be spelled out purely in terms of the statistical properties of those methods, without reference to facts about how those methods work or the surrounding sociohistorical context.

---

[1] Angwin et al. (2016).
[2] These formulations of Equal False-Positive Rates and Equal False-Negative Rates are from Hedden (2021).
[3] Hardt et al. (2016).

In this paper, I propose a novel version of Equalized Odds and argue that it is both intuitively plausible and avoids three key problems for the criterion. The first problem, which I will call *the problem of ideal accuracy*, is that even an infallibly accurate predictive method can violate Equalized Odds when base rates of the feature being predicted differ across groups. I argue that this problem does not show that Equalized Odds is a requirement of fairness, but instead teaches us something important about how Equalized Odds should be understood.

The second problem comes from Hedden (2021), who has recently argued that Equalized Odds is not a necessary condition of fairness by appealing to a counterexample featuring people in two different rooms flipping coins of varying weights.[4] In the example, being in one of the two rooms is supposed to be analogous to belonging to one of two social groups, and the weight of a person's coin is supposed to correspond to their risk level of a particular kind, such as their risk of recidivating or defaulting on a loan. Hedden appeals to the case to argue that a perfectly fair algorithm will nonetheless violate Equalized Odds under some conditions. This seems to show that Equalized Odds is not a requirement of fairness.

The third problem appeals to another popular statistical criterion of fairness, Calibration Within Groups:

> Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group.[5]

It can be shown that Calibration Within Groups is incompatible with Equalized Odds in cases where base rates of the feature being predicted differ across groups. Some authors take this to suggest that Equalized Odds is not a requirement of fairness.[6] While a full treatment of this problem is outside the scope of this paper, I provide a preliminary response in the conclusion.

The plan is as follows. In section 2, I set out and motivate my version of Equalized Odds. I also argue that Equalized Odds can explain the same case judgments as Base Rate Tracking, a statistical criterion of fairness proposed by Eva (2022), and raise an objection to the latter criterion. In section 3, I address the problem of ideal accuracy, and show that it arises from a misunderstanding of how to apply Equalized Odds to cases where objective chanciness is involved. That response commits my interpretation of Equalized Odds to nontrivial metaphysical assumptions

[4] See Flores et al. (2016), Huq (2019), Mayson (2019), Hellman (2020), and Long (2021) for further objections that are outside the scope of this paper.

[5] This formulation of Calibration Within Groups is from Hedden (2021), 214.

[6] See e.g. Dieterich et al. (2016), Corbett-Davies et al. (2016), and Long (2021).

about objective chance; section 4 argues that those assumptions are reasonable. Section 5 considers Hedden's objection as it applies to my version of Equalized Odds, and argues that it can be answered by taking Equalized Odds to be a pro tanto requirement of procedural fairness rather than an all-things-considered requirement. Section 6 briefly addresses the conflict between Equalized Odds and Calibration Within Groups and offers concluding remarks.

2 *Equalized Odds*

In this section, I explain and motivate my proposed version of Equalized Odds. As I will understand it, Equalized Odds applies to a class of decision problems that I will call "qualification problems." These are cases where a decision-maker must decide whether to allocate some benefit or burden to particular individuals on the basis of whether they possess some feature that morally justifies allocating that benefit or burden to them. Candidate examples of qualification problems include deciding whether

1. a defendant in a criminal trial is innocent, and so ought to be acquitted;
2. a patient has a deadly and highly transmissible disease, and so ought to be quarantined (or receive a highly effective but scarce treatment);
3. an adult is the biological parent of a child, and so ought to be granted visitation rights (or required to pay child support);
4. an insurance subscriber has filed a valid claim, and so ought to be reimbursed;
5. a defendant in a criminal trial is likely to recidivate, and so ought to be detained pretrial;
6. a patient has a mental illness that presently renders them a danger to themselves or others, and so ought to be involuntarily committed;
7. a loan applicant is likely to repay a loan, and so ought to have their application approved; or
8. a parent is at high risk of neglecting or abusing their child, and so ought to lose custody of the child.

The distinctive features of qualification problems are as follows. First, an institutional decision-maker is deciding which individuals in some group of decision subjects to allocate some benefit or burden to. Second, there is some feature of individuals whose presence or absence determines whether allocating the benefit or burden to them would be substantively fair. For example, if a defendant in a criminal trial is *innocent of the crime*

they are accused of, then the substantively fair result is a finding of innocence. Call the features that substantively justify allocating the relevant benefit or burden *qualifications*, and call an individual *qualified* when they have those features. Third, since the relevant qualifications cannot be directly observed, they must be inferred from the available evidence. Some errors in determining whether an individual is qualified are thus inevitable, making qualification problems instances of what Rawls called "imperfect procedural fairness"[7]and giving rise to claims of procedural fairness regarding how those errors tend to be distributed across different groups of decision subjects.

As formulated above, Equalized Odds makes reference to individuals being "actually positive" and "actually negative," as well as to individuals being "falsely predicted to be positive" and "falsely predicted to be negative." I will understand these terms in the following way, as applied to decision procedures. Say that a decision subject is an "actual positive" just in case they are qualified (in the sense defined above), and an actual negative otherwise. For example, a defendant in a criminal trial is an actual positive iff they are innocent. Further, say that a subject is "falsely predicted to be positive" when they are not qualified but are incorrectly judged to be qualified by the institution in question, and say that a subject is "falsely predicted to be negative" when they are qualified but incorrectly judged not to be qualified.

We can then restate Equalized Odds as follows:

> (1) The expected percentage of actually unqualified individuals who are falsely judged to be qualified is the same for each relevant group *and* (2) the expected percentage of actually qualified individuals who are falsely judged to be unqualified is the same for each relevant group.[8]

My claim is that Equalized Odds, so understood, is a requirement of procedural fairness. If Equalized Odds is not satisfied by a decision

---

[7] Rawls (1999), 74–75.

[8] The probabilistic expectations here should be understood in an externalist way, in terms of objective chances. Hedden (2021) suggests that understanding them in terms of objective chances will only work when there is "objective chanciness involved" (see fn. 15). However, we can (and I think should) always understand the relevant probabilities in terms of objective chances. On my interpretation of Equalized Odds, whether a given decision subject is qualified is not chancy but rather determinate at the time of decision (see section 3). There will, however, be objective chanciness in how the decision procedure being evaluated classifies particular decision subjects. This will be true even if the decision procedure consists of a deterministic algorithm; such an algorithm must be implemented on physical hardware that might malfunction, in addition to chanciness introduced by how input data is collected and processed.

procedure, then using that procedure to make decisions of the relevant kind would be unfair to members of the group that are thereby disadvantaged.

Equalized Odds is normally understood as a constraint on binary classifiers. A binary classifier is a predictive algorithm that attempts to classify individuals as belonging to one of two categories on the basis of data about their other features. Spam filters, for example, attempt to classify email messages as "spam" or "not spam" on the basis of features such as the identity of the sender and the content of the message. Rather than understanding Equalized Odds to apply directly to binary classifiers, I will instead interpret it as applying to decision procedures used to solve qualification problems. For example, some criminal courts in the United States use the following decision procedure to determine whether a defendant should be granted pretrial bail: the judge presiding over the pretrial hearing is provided with information about the defendant's criminal history as well as a risk score generated by a recidivism prediction algorithm such as COMPAS, attempts to discern whether the defendant presents a sufficiently grave risk to the public to justify pretrial detention, and then grants or denies bail on the basis of their professional judgment. As I understand Equalized Odds, it applies to this procedure taken as a whole, rather than merely to the algorithm that supports the judge's decision-making.

In understanding Equalized Odds as applying to decision procedures rather than predictive methods, I am diverging from some of the recent literature on statistical criteria of fairness. Both Eva and Hedden focus on the fairness of *predictive methods*, rather than with the fairness of *decision procedures*. I am focusing on the latter for two reasons.

First, it is not clear that predictive methods can be fair or unfair considered in themselves. Consider recidivism prediction algorithms, for instance. Suppose a recidivism prediction algorithm tends to overestimate recidivism risk in the case of black defendants, but underestimate it in the case of white ones. Is the algorithm unfair to black defendants? Our judgments here seem to depend on the surrounding institutional context. On the one hand, if the algorithm is used to decide whether to grant bail, then being rated as high risk is a bad thing, and intuition suggests that the algorithm's predictions treat black defendants unfairly. On the other hand, if the algorithm is used to decide whether to provide targeted assistance that reduces recidivism, such as free subsidized housing or counseling, then intuition suggests that the algorithm is instead unfair to white decision subjects. Alternatively, suppose that researchers build an insurance fraud detection algorithm solely for purposes of studying its mathematical properties, but test it on data from real loan applicants. Suppose also that the algorithm generates false positives for poor claimants at such high rates that it would be obviously be unfair to be used as a basis for denying claims. If the researchers know that the algorithm

will never be used, are poor claimants represented in training data nonetheless treated unfairly by the algorithm's predictions? It seems to me that they only have cause for complaint about the algorithm's behavior if decisions are subsequently based on them—or at any rate might be.[9] These considerations suggest that the fairness of a predictive method cannot be judged independently from how its predictions are used, and are better understood as judgments about the fairness of the decision procedure it is embedded in.

Second, even if we can make sense of the idea that predictive methods are fair or unfair considered in themselves, it nonetheless makes more sense to focus on the fairness of decision procedures in the present context because it is the kind of fairness that is at issue in policy debates about statistical criteria of fairness. When ProPublica claimed that COMPAS was unfairly biased against black defendants, for instance, this was presumably shorthand for the claim that the practice of *using* COMPAS to make certain *decisions* about how to treat black defendants is unfair. After all, if COMPAS were not used to make important decisions about how to treat black defendants, then it would not be clear that the supposed bias ProPublica identified would be a matter of public concern. Their complaint is thus best understood as a complaint of procedural fairness, not predictive fairness considered in isolation from procedural fairness.

Even though Equalized Odds applies in the first instance to decision procedures rather than the predictive algorithms they employ, it nonetheless has important implications for how the latter ought to be designed. The reason for this is that whether a decision procedure satisfies Equalized Odds will depend largely on the fact-finding methods it uses to assess whether particular decision subjects are qualified. If Equalized Odds is a requirement of procedural fairness for qualification problems, then the designers of predictive algorithms used to solve qualification problems have a duty to design their algorithms in a way that will tend to result in Equalized Odds being satisfied.

Why, intuitively, is Equalized Odds a requirement of procedural fairness in qualification problems like (1)–(8) above? Suppose that Equalized Odds is violated in criminal trials. This would mean that there are two social groups A and B such that either (1) innocent members of A are more likely to be convicted than innocent members of B, or (2) guilty members of A are more likely to be acquitted than guilty members of B. Suppose, for example, that black defendants in the United States are more

[9] Perhaps poor claimants are wronged if the researchers *believe* the algorithm's predictions. As Hedden points out (p. 220), defenders of moral encroachment argue that we can wrong others simply by believing certain things about them; see e.g. Moss (2018) and Basu (2019a, 2019b). However, whether anyone believes an algorithm's predictions is not something that is intrinsic to the algorithm itself.

likely to be mistakenly convicted than white defendants. Intuitively, this constitutes a procedural injustice against black defendants (or at any rate innocent ones).[10] Similarly, suppose that guilty white defendants are more likely to be acquitted than guilty black defendants. This too, would seem to constitute a procedural injustice against black defendants.[11] Similar points apply to the other cases mentioned above. This suggests that Equalized Odds is a requirement of procedural fairness in a wide range of qualification problems.

Importantly, Equalized Odds can also explain our judgments in a class of cases stressed by Benjamin Eva. In the example Eva considers, we consider a credit scoring algorithm that assigns risk scores to decision subjects based on their zip code as described in the following table:[12]

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	10%	25%
White	TR10	Bad	30	20%	25%
White	TR11	Good	40	10%	75%
White	TR11	Bad	40	20%	75%
Black	TR10	Good	60	10%	25%
Black	TR10	Bad	20	20%	25%
Black	TR11	Good	60	10%	75%
Black	TR11	Bad	60	20%	75%

The algorithm produces these scores by "redlining" decision subjects based on zip code: residents of the majority black zip code (TR11) receive a risk score of 75% whereas residents of the majority white zip code (TR10) receive a risk score of 25%. In so doing, the algorithm ignores available information about credit risk. In particular, Eva stipulates that credit score tracks the probability that a given resident will default perfectly—residents with good credit default 10% of the time, whereas residents with bad credit default 20% of the time. Eva argues, plausibly, that "[b]y ignoring credit score and basing risk scores purely on applicants' zip

[10] As Di Bello and O'Neil (2020) point out, and take to motivate a criterion of procedural justice in criminal trials that they call "equal protection," which requires that "innocent defendants not be exposed to higher risks of mistaken conviction than other innocent defendants facing the same charges or comparably serious charges" (158).

[11] Intuitions in favor of Equalized Odds are especially strong when the disadvantaged group more socially marginalized than the advantaged group. See Castro (2019) for one possible explanation of the asymmetry.

[12] Table reproduced from Eva (2022), 254. I have substituted percentages for fractions for convenience.

codes, the algorithm seems to treat black applicants unfairly in comparison to white applicants."

Eva takes this example to motivate a novel statistical criterion of fairness, Base Rate Tracking:

> Base Rate Tracking: The difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups.[13]

According to Eva, Base Rate Tracking explains why the algorithm in question is unfair:

> Note first that the overall average risk score for white applicants is 9/20, while the overall average risk score for black applicants is 11/20. Next, note that the overall default rate for white applicants is 27/200, while the overall default rate for black applicants is 28/200. So while the difference between the average risk scores of white and black applicants is 2/20, the difference between the overall default rates of white and black applicants is only 1/200. The difference between the average risk scores of the two groups is 20 times as great as the difference between their actual default rates. This, it seems to me, is a clear indication of unfairness. If an algorithm assigns one group a higher average risk score than another, that discrepancy has to be justified by a corresponding discrepancy between the base rates of those two groups, and the magnitudes of those discrepancies should be equivalent.[14]

I agree with Eva that the way the algorithm assigns risk scores seems unfair. However, we do not need to appeal to Base Rate Tracking to explain *why* it seems unfair. In my view, Equalized Odds provides a more natural explanation. Eva stipulates that the algorithm in question was developed by a bank in order to decide which loan applications to accept. Since the algorithm produces only two risk scores, 25% and 75%, the only reasonable assumption is that the bank plans to approve loans from applicants with the former risk score only. Further, Eva has stipulated that there are only two kinds of applicants, those with good

[13] Eva (2022), 258.

[14] Eva (2022), 258. Note that Eva concedes that there is an alternative way to explain why the algorithm's predictions are unfair: one might think that it is unfair to base lending decisions on zip code because "the correlations between race, zip code and default rates are themselves the product of unjust social economic historical trends" (p. 255). However, Eva maintains that "there is something intrinsically unfair in the predictions themselves, [and] we should not need to refer to the predictive features used by the algorithm in order to diagnose that unfairness. [W]e should be able to diagnose the intrinsic unfairness of the algorithm's predictions using statistical criteria alone" (p. 257).

credit and those with bad credit, and that credit score is a "perfect indicator" of residents' "true" default risk. This suggests that applicants qualify for loans just in case they have good credit rather than bad credit.

Once these implicit assumptions are made explicit, it becomes clear that the decision procedure the bank plans to use to decide which loan applications to approve violates Equalized Odds. The expected percentage of qualified white applicants who will mistakenly be denied a loan is about 31%, compared with 50% of qualified black applicants. Similarly, the expected percentage of unqualified white applicants who will mistakenly receive a loan is approximately 43%, compared with only 25% of black applicants. Therefore, both parts of Equalized Odds are violated in a way that is intuitively unfair to black applicants.[15]

Further, Base Rate Tracking fails to yield the right result in cases where it comes apart from Equalized Odds. Consider a credit scoring algorithm that violates Base Rate Tracking for the following reason: for scores that are well below the bank's decision threshold, it overestimates default risk for black applicants relative to white applicants (but not nearly enough to put them in danger of being rejected). For all other scores, the algorithm neither over- nor underestimates the default risk of black applicants relative to white applicants. Further, suppose the bank is keen to avoid treating black applicants unfairly, but continues using the algorithm because it knows that the algorithm is biased in a way that will not affect anyone's chances of being approved. In my view, there is nothing unfair about the bank continuing to use the algorithm. This shows that Base Rate Tracking is not a requirement of procedural fairness, and suggests that the cases that seem to support it instead motivate Equalized Odds.[16]

3 The problem of ideal accuracy

I said at the outset that ProPublica's argument against COMPAS implicitly appealed to a version of Equalized Odds. In fact, it is natural to interpret

[15] While the case judgments I have discussed support Equalized Odds, a full defense would require investigating *why* it is a requirement of fairness, which is outside the scope of this paper. Castro (2019) develops one possible argument, focusing on the special case of recidivism prediction.

[16] Note that Eva's focus is on what they call "intrinsic fairness," which concerns the fairness of an algorithm's predictions considered in isolation from how they are used as well as other features of the surrounding social and historical context. It is open to Eva to accept that Base Rate Tracking is not a requirement of procedural fairness, but maintain that it is nonetheless a criterion of intrinsic fairness. I submit, however, that there is nothing unfair about the situation just described, which suggests that Base Rate Tracking is not a criterion of intrinsic fairness, either.

ProPublica as tacitly appealing to *my* version of Equalized Odds. Consider the crucial passage from their exposé:

> "[We] turned up significant racial disparities .... In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.
>
> The formula was particularly likely to *falsely flag black defendants as future criminals*, wrongly labeling them this way at almost twice the rate as white defendants" (emphasis mine).[17]

With this passage in mind, here is what I take to be the most natural way to reconstruct ProPublica's argument that COMPAS is unfairly biased against black defendants.[18] Statistical analysis shows two things about COMPAS: it has a higher false-positive rate for black defendants than white defendants, and a higher false-negative rate for white defendants than black defendants. The fact that COMPAS' false-positive rate is higher for black defendants entails that COMPAS is more likely to falsely judge that a defendant will commit a future crime if the defendant is black than if the defendant is white. Similarly, the fact that COMPAS false-negative rate is higher for white defendants entails that COMPAS is more likely to falsely judge that a defendant will *not* commit a future crime if the defendant is white than if the defendant is black. If judges base their pretrial detention decisions on COMPAS scores, then their judgments about which defendants qualify for pretrial detention will exhibit a similar pattern of errors. Intuitively, this would be unfair to black defendants, because it would violate Equalized Odds as defined above. Using COMPAS to make pretrial detention decisions is therefore unfair.

In response to this argument, Equivant's researchers accused ProPublica's analysis of a variety of methodological deficiencies.[19] One of the most important of these was that false-positive and false-negative rates are unreliable indicators of bias in cases where base rates of the feature being predicted differ across groups:

> Results of our analyses indicate that as the mean difference in scores between a low-scoring group and a high-scoring group is increased, the base rates diverge and higher false positive rates and lower false negative rates are obtained for the high-scoring group. This is the same pattern of results reported by Angwin et al. This pattern does not show evidence of bias, but rather is a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores.[20]

[17] Angwin et al. (2016).

[18] See Castro (2019) for an alternative reconstruction of ProPublica's argument.

[19] Dietrich et al. (2016).

[20] Dietrich et al. (2016), 8.

Equivant's researchers make two claims in this passage. First, disparities in base rates of the feature being predicted will normally produce disparities in false-positive and -negative rates—even if an entirely unbiased predictive method is used. Second, it follows that disparities in false-positive and -negative rates do not provide evidence that a predictive method is biased in cases where base rates diverge, because we would expect to see such disparities *regardless* of whether the method is biased.

Given that the black defendants in ProPublica's dataset do appear to have experienced a higher base rate of recidivism than their white counterparts, these claims directly undermine ProPublica's argument if true.[21] More importantly for our purposes, they seem to pose a serious challenge to Equalized Odds. If there are cases in which (a) a decision-maker uses an entirely unbiased method to determine whether decision-subjects are qualified and (b) the decision-maker nonetheless violates Equalized Odds, then doesn't that show that violations of Equalized Odds are not always procedurally unfair? If so, then Equalized Odds cannot be a requirement of procedural fairness.

In this section, I will develop what I take to be the strongest version of this argument. My version will focus on a procedure for estimating recidivism risk that everyone should concede is not unfairly biased on the basis of race: a decision procedure whose estimates of recidivism risk are *infallible*, in the sense that they never classify a high-risk defendant as at low-risk of recidivism or vice versa. I will show that even a decision procedure that is infallible in this sense can nonetheless produce false-positive and -negative rate disparities of the kind that ProPublica observed and took to show that COMPAS is unfairly biased. Call this the *problem of ideal accuracy*.

The problem of ideal accuracy initially seems to show that Equalized Odds is not a requirement of fairness. However, I will argue that it instead shows that more care needs to be taken when we apply Equalized Odds to qualification problems where objective chances are in play.

To see that even an infallible procedure for estimating recidivism risk can nonetheless produce significant false-positive and -negative rate disparities, consider the following case, *Jewel Thieves*:

> Suppose that there are two guilds of jewel thieves, the Professionals and the Hobbyists. Each thief has made the

[21] Note that ProPublica used *being charged with a new crime* as a proxy for recidivism, as is standard in the field of recidivism prediction. Given racial disparities in policing, one might worry that this practice will tend to result in decision-makers overestimating the recidivism risk posed by black defendants relative to white defendants (see Mayson 2019 and Long 2021 for discussion).

> following deal with God: God will use a random number generator between 0 and 100 to determine whether they go on to steal (by either ensuring they will steal successfully, or frustrating their efforts). The terms of the deal, however, are different for Professionals and Hobbyists. If the thief is a Professional, the deal is that God will induce them to steal with a 90% probability and prevent them from stealing with a 10% probability. For Hobbyists, these odds are reversed. God will ensure that there is a 10% probability that each Hobbyist will steal, and a 90% probability that they won't. This is admittedly fanciful, but it's a straightforward way to ensure that the objective chance that each Professional will steal is 90%, and the objective chance that each Hobbyist will steal is 10%. Suppose also that all of the thieves wear hats, which are either blue or green. Among the Blue Hats, 100 are Professionals and only 10 are Hobbyists. Among the Green Hats, those numbers are reversed; only 10 are Professionals and 100 are Hobbyists.
>
> Now suppose that you are a pretrial hearing judge, and all of these thieves are going to appear before you in court. Fortunately, you have a perfectly reliable way to estimate recidivism risk—God will helpfully tell you the objective chance that each thief will steal within two years, based on whether they are a Professional or a Hobbyist. Assume for concreteness that a defendant poses a sufficient threat to the public to make preventive detention (objectively) justified if the objective chance that they will steal in the future exceeds 50%. Knowing this, you adopt the following policy: if God tells you that the objective chance that a given thief will steal exceeds 50%, you will classify them as "high risk" and detain them; otherwise, you will classify them as "low risk" and release them.

ProPublica's analysis of COMPAS defined "false-positive" and "false-negative" in the following way:

> A defendant is a *false-positive* iff they (1) were labeled "high risk" and (2) did not recidivate.
>
> A defendant is a *false-negative* iff they (1) were labeled "low risk" and (2) did recidivate.

Applying these definitions to Jewel Thieves yields the following false-positive and -negative rates for the Blue Hats and Green Hats:

	Professionals	Hobbyists	FP	AN	FPR	FN	AP	FNR
Blue Hats	100	10	10	19	~53%	1	91	~1%
Green Hats	10	100	1	91	~1%	10	19	~53%

As you can see, the false-positive rate for the Blue Hats is more than fifty times the false-positive rate for the Green Hats, and the false-negative rate for the Green Hats is more than fifty times the false-negative rate for the Blue Hats! These are exactly the sort of findings that led ProPublica to conclude that using COMPAS to make pretrial detention decisions is unfair to black defendants.

Something, however, has clearly gone wrong. Despite the dramatic disparities in false-positive and false-negative rates, it is clear that your decision procedure is not unfairly biased against the Blue Hats. After all, your method for estimating recidivism risk is infallible—God just tells you each thief's objective chance of stealing—which means that you never misclassify anyone who is objectively low risk as high risk. All of the Professionals are correctly classified as high risk, since their objective chance of stealing at the time hearings occur exceeds 50%. And all of the Hobbyists are correctly classified as low risk, since their objective chance of stealing does not exceed 50%. Detaining all of the Professionals and releasing all of the Hobbyists thus seems like the substantively fair result. In light of this, the disparate false-positive and false-negative rates do not provide grounds for thinking that your decision procedure is unfairly biased against the Blue Hats. This seems like a straightforward counterexample to Equalized Odds.

Jewel Thieves thus suggests that Equivant was right: disparities in false-positive and -negative rates are not a reliable indicator of unfair predictive bias. Should we conclude that Equalized Odds is not a requirement of procedural fairness? I do not think that we should, because ProPublica's way of understanding false-positive and false-negative rates is not the right one given the intuitive motivation for Equalized Odds.

As discussed above, the motivation for my version of Equalized Odds assumes a particular way of understanding what it means to be a "false-positive" and "false-negative," where being "positive" is understood in terms of being qualified for more favorable treatment. To determine how Equalized Odds should be applied to any particular case, then, we need to first determine what features justify allocating the benefit or burden in question to particular individuals, rendering their receiving that burden/benefit substantively fair.

Qualification problems come in two flavors. On the one hand, say that a qualification problem concerns qualifications that are *dispositional* just in case a decision subject is qualified iff their objective chance of coming to have feature F at some future point exceeds some decision threshold. On the other hand, say that qualifications are *categorical* when

a decision subject is qualified iff they now have feature F, and having feature F is not a matter of one's objective chances to have some other feature or features. In cases (1)–(4) the qualifications are categorical: the defendant either committed the crime or didn't, the patient either has the disease or doesn't, the adult is either the biological parent or isn't, and the subscriber either has a valid claim or doesn't. By contrast in cases (5)–(8) the qualifications appear to be dispositional: there is a chance that the defendant will commit a crime, a chance that the patient will harm themselves or others, a chance the loan applicant will default, and a chance the parent will abuse or neglect their child. In these cases there is some probabilistic threshold (which may vary across individuals and groups[22]) above which the decision subject qualifies to receive the relevant benefit or burden; a decision subject ought to receive that benefit or burden if and only if they exceed that threshold.

ProPublica appears to have assumed that a defendant in a pretrial detention hearing is qualified to receive a burden—pretrial detention—just in case they *subsequently commit a crime within two years*. On this understanding, a defendant is a "false positive" just in case they (1) are classified as a future recidivator but (2) do not recidivate. This understanding assumes that the qualifications that are relevant when judges base pretrial detention decisions on recidivism risk are *categorical*, in the sense just defined. There is, however, an alternative way to think about when defendants qualify for pretrial detention. We might instead assume that the relevant qualifications are *dispositional*—i.e., that defendants qualify for pretrial detention on the basis of recidivism risk just in case they are *at objectively high risk of recidivism*.[23] On this understanding, a defendant is a "false positive" just in case they (1) are classified as at high risk of recidivism, but (2) do not recidivate.

This ambiguity in how we should understand what it means to be a "false positive" or a "false negative" seems to have passed unnoticed in the debate about whether Equalized Odds is a requirement of fairness. Which understanding is correct depends on when defendants actually qualify for pretrial detention—that is, on when detaining a defendant pretrial is substantively fair. Does whether detention is substantively fair depend on whether the defendant is at objectively high risk of recidivism, or whether they will actually recidivate?

In my view, the former claim is far more plausible. To see this, consider the following case, *Sutton's Conversion*:

---

[22] Various authors have suggested that the appropriate threshold might be different for different individuals. See e.g. Castro (2019), Huq (2019), and Long (2021). I set aside the difficult question of what kinds of facts determine what threshold is appropriate, as well as the complication that other factors might be relevant to how decision subjects ought to be treated.

[23] To simplify discussion, I am assuming that the practice of pretrial detention on the basis of recidivism risk is morally justifiable.

> The famous bank robber Willie Sutton is once again on trial for (you guessed it) bank robbery. During his pretrial hearing, Sutton is understandably classified as at high risk of recidivism by COMPAS, and ordered detained by the judge. However, Sutton is accidentally released, and the charges against him are dropped. While on his way to rob the local bank, Sutton is nearly killed when a truck runs a red light. As a result of this near-death experience, Sutton experiences an (antecedently very unlikely) religious conversion and never robs another bank.

We can safely stipulate that, at the time of his hearing, Sutton's objective chance of recidivism was extremely high.[24] The question is whether the substantively fair result was for Sutton to have been detained pretrial, given that he did not subsequently recidivate. In my view, the answer is a clear "yes." Given that Sutton was in fact at extremely high risk of recidivism at the time of his hearing, the court made the correct decision in deciding to detain him: the fact that an unlikely chance event resulted in his antecedently high *risk* of recidivism not manifesting in *actual* recidivism does not suggest that the court's decision to detain him was substantively unfair.

Moreover, the alternative understanding of when defendants qualify for pretrial release generates implausible results in Jewel Thieves. Suppose that pretrial detention is substantively fair just in case the defendant will in fact commit a future crime. It follows that the Blue Hats are far more likely than the Green Hats to be treated in ways that are substantively unfair. But if so, it is hard to avoid the conclusion that there is procedural unfairness in Jewel Thieves. Isn't there *something* procedurally unfair about subjecting the Blue Hats to a far greater risk of mistaken detention than the Green Hats? But intuitively, there is *no* sense in which the Blue Hats are being treated less fairly than the Green Hats. Something has gone wrong, namely the assumption that detention is substantively fair just in case the defendant will actually recidivate in the future.

By contrast, if we drop this assumption in favor of the proposed alternative—detention is substantively fair just in case the defendant's objective chance of recidivism is sufficiently high—we get the intuitively correct result in Jewel Thieves. Recall that we assumed above that detention is substantively fair iff the defendant's objective chance of recidivism exceeds 50%. A defendant in Jewel Thieves is a false positive

[24] Sutton wrote the following in his autobiography: "Why did I rob banks? Because I enjoyed it. I loved it. I was more alive when I was inside a bank, robbing it, than at any other time in my life. I enjoyed everything about it so much that one or two weeks later I'd be out looking for the next job" (Sutton and Lynn 2004).

in the sense that is relevant to my version of Equalized Odds, then, just in case (1) the objective chance that they will recidivate (steal) is below 50% but (2) they are misclassified as having an objective chance of recidivism in excess of 50%, and so misclassified as unqualified for pretrial release. Further, the false-positive rate (in the relevant sense) is the probability that a randomly selected defendant whose risk of recidivism is below 50% will be incorrectly classified as at high risk of recidivism and detained. Using this definition of false-positive rates, the false-positive rate for both Blue Hats and Green Hats is 0%: since our method for estimating recidivism risk is perfectly accurate, defendants whose objective recidivism risk is low are correctly classified as low risk 100% of the time. The false-negative rate for both groups is also 0%, again using the appropriate definition. Jewel Thieves, then, is not a case in which a perfectly fair predictive method violates my version of Equalized Odds, and so not a counterexample to my claim that Equalized Odds is a requirement of procedural fairness.

As we have seen, even an infallibly accurate method for assessing whether decision subjects are qualified can generate false-positive and false-negative rates that vary significantly across groups provided the risk distributions also differ for those groups. But this is only true if we understand false-positive and false-negative rates as ProPublica did in conducting their analysis of COMPAS. As I have shown, though, this is not the right way to understand false-positive and false-negative rates for purposes of applying Equalized Odds in cases where the relevant qualifications are dispositional. The problem of ideal accuracy therefore fails to show that Equalized Odds is not a requirement of procedural fairness.

4 Worries about objective chances

In responding to the problem of ideal accuracy, I made the following assumption about the normative structure of decision problems like (5)–(8) above: whether it would be substantively fair for a given decision subject to receive favorable treatment is determined by whether their objective chance of engaging in the relevant behavior exceeds some threshold. I also assumed that the relevant objective chances are not simply determined by whether the decision subject *actually* engages in the relevant behavior, in which case a defendant's risk of recidivism would be 100% if they will reoffend and 0% if they will not. Instead, I assumed that decision subjects typically have *nondegenerate* objective chances of doing various things in the future, such as committing crimes or repaying loans.

There are two worries one might have about these assumptions. First, one might worry that they are problematic outside of fanciful cases like Jewel Thieves, where suitable objective chances are simply stipulated into existence. What could it possibly mean, for example, to say that a real-

world defendant has an objective chance of committing a crime within two years that is above 50%? Second, *even if* we assumed that objective chances of the requisite kind exist, one worry that they would not be measurable. This is a problem on the plausible assumption that something cannot be a requirement of procedural fairness if there is no way for decision makers to tell whether they are satisfying it. Unlike substantive fairness, procedural fairness must be accessible to us to some degree.[25]

The conception of objective chance that I have in mind is familiar from both everyday life and scientific theory. We ordinarily assume that some people are more likely to experience particular outcomes—behavioral or otherwise—than others. We assume that young children are more likely to cry in grocery stores than grown adults. We assume that smokers are more likely to develop cancer than nonsmokers. We assume that professional baseball players are more likely to hit home runs than Supreme Court justices. We assume that first-generation students are more likely to struggle in our classes without additional support than students of college professors.

Further, we assume that the "likelihoods" just mentioned are objective in some sense, rather than being claims about what is reasonable to believe. For example, the claim that smokers are more likely to develop lung cancer is not a claim about what a reasonable person would believe given the available evidence, but about the causal link between smoking and lung cancer. And we assume that have various more-or-less reliable ways of estimating the objective chances of various outcomes. Informal observation and common sense tells us that a professional baseball players is unlikely to hit a home run at any given chance at bat, but still much more likely than a Supreme Court justice. Through scientific investigation, we can give far more precise estimates of individuals' objective chances of particular outcomes, such as estimates of the chance that a patient with pneumonia will die within 30 days based on various risk factors. In other words, we assume that individual's objective chances of various outcomes are often measurable informally to some extent, and that we can improve our informal estimates by conducting more rigorous research.[26]

Finally, we assume that how we ought to treat people very often depends on their objective chances of experiencing particular outcomes,

---

[25] Note that, if Equalized Odds is a requirement of procedural fairness, then it follows that procedural fairness is not perfectly accessible to us, as Equalized Odds is an externalist constraint. Other authors have endorsed the idea that there are externalist constraints on procedural fairness; see e.g. Gardiner (2019).

[26] Strevens (1999) notes that "probabilistic generalization is the rule in the medical sciences" (244), adding that the relevant probabilities should be understood objectively. Just how we should understand objective chances of macroscopic events is a vexed issue, and beyond the scope of this paper. Various accounts of objective chance are available that are compatible with my defense of Equalized Odds; see for example List and Pivato (2015) and Glynn (2010).

and that we are obligated to seek out evidence that will help us estimate those chances. For example, patients who are likely to die of pneumonia if released ought to be admitted to the hospital. Importantly, this isn't just a claim about subjective chance. While it's true that patients that are *subjectively* likely ought to be admitted to the hospital, it's also true that patients that are *objectively* likely ought to be admitted to the hospital. The latter claim has implications that the former claim does not. For one thing, it helps explain why a doctor deciding whether to admit a patient with pneumonia ought to first gather additional evidence of the patient's pneumonia mortality risk, such as by measuring the patient's blood urea nitrogen level.[27] For another, it helps explain why hospitals ought to invest in updating their methods for estimating pneumonia risk over time, thereby making new kinds of evidence available to doctors and putting them in a better position to assess mortality risk. The same holds when the outcomes in question are behavioral, as is the case in many of the dispositional qualification problems I have mentioned. For example, psychiatric inpatients who are likely to commit suicide ought to be monitored closely.

Let's return to the special case of recidivism risk. Why should we think that there are nondegenerate objective chances that particular defendants will commit new crimes, and that they we are in a position to measure those chances to some extent?

First, the claim that defendants have (nondegenerate) objective chances of recidivating is difficult to deny, at least from a pretheoretical perspective. It is overwhelmingly natural to assume that different people have different tendencies to commit crimes, due to a combination of what they are like as well as what their environment is like, and that this means that criminal defendants will vary in how (objectively) likely it is that they will recidivate. I believe (and imagine you do, too) that Joe Biden's tendency to rob banks is negligible, and that the probability that Joe Biden will rob a bank within two years is extremely low. By contrast, when the famous bank robber Willie Sutton was in his prime, he was strongly disposed to rob banks, a tendency that was partly the product of his great passion for robbing them and partly due to their availability in Sutton's environment. As a consequence, the likelihood that Sutton would rob a bank within two years was quite high. Further, I believe (and imagine you do too) that these claims are objective in the sense that they are made true by facts about Biden and Sutton and their respective environments, not by facts about what it is rational for us to believe about Biden and Sutton given the available evidence.

Second, it is also reasonable to assume that objective chances of recidivism are measurable to a significant extent. The explicit goal of the

[27] Cf. Smith (2014), who argues that subjective moral theories cannot explain the duty to gather evidence before acting.

researchers who develop recidivism prediction instruments is to discriminate between individuals who are likely to commit future crimes and individuals who are not. They attempt to do this by constructing statistical models based on observed frequencies of recidivism among known offenders in conjunction with background knowledge about the causes and correlates of recidivism from decades of criminological research. That they are able to discriminate high-risk and low-risk defendants fairly successfully is demonstrated by the fact that higher risk scores have a strong correlation with observed frequencies of recidivism. For example, an independent analysis of data collected in Broward County, FL found that defendants that received the highest COMPAS score reoffended about 81 percent of the time, whereas defendants that received the lowest score reoffended about 22 percent of the time.[28] This gives us evidence that COMPAS scores measure defendants' objective chances of committing crimes with a significant degree of accuracy.

Summing up, the assumption that defendants have nondegenerate objective chances of recidivism and that these chances are measured by recidivism prediction instruments, albeit imperfectly, is pretheoretically reasonable. Indeed, it seems hard to deny. It is also consistent with a variety of accounts of the nature of objective chance (see footnote 26). Similar remarks apply to the objective chances featured in other qualification problems, such as those listed in section 2 above. While my defense of Equalized Odds carries nontrivial assumptions about objective chances, it is reasonable for us to accept those assumptions absent a compelling argument to the contrary.

5 The problem of infra-marginality

Brian Hedden has recently argued that Equalized Odds is not a requirement of procedural fairness by appeal to what statisticians call *the problem of infra-marginality*.[29] The problem of infra-marginality was first introduced by Ayres (2002) as an objection to outcome-based tests for taste-based discrimination. In taste-based discrimination, a decision-maker treats members of a social group less favorably than others because she prefers to treat them less favorably (e.g., because of animus), as opposed to treating them less favorably because the available evidence suggests they are less qualified than others (Becker 1957). Outcome-based tests for taste-based discrimination attempt to detect taste-based discrimination by observing the "success rate" for decisions affecting members of different social groups (Becker 1957, 1993). For example, an outcome-based test for taste-based discrimination in vehicle searches by

[28] Corbett-Davies et al. (2016).

[29] I am grateful to an anonymous reviewer for suggesting that I address this objection to Equalized Odds more explicitly.

police might compare the rate at which such searches find contraband for white and black motorists. If searches of white motorists' vehicles discover contraband more often than searches of black motorists' vehicles, then this is treated as evidence that police are engaging in taste-based discrimination against black motorists (Simoui et al. 2017).

The assumption that motivates outcome-based tests is that if a decision-maker D's success rate is higher for group A than group B, then it must be the case that D is using a higher "decision threshold" for As than Bs, in the sense of requiring there to be more evidence that a given decision subject is qualified to receive the relevant benefit or burden when that decision subject is an A. This would seem to constitute taste-based discrimination against either As (if a benefit is being allocated) or Bs (if a burden is being allocated).[30]

As Ayres pointed out, the problem with this reasoning is that D's success rate for groups A and B does not depend solely on the decision-threshold that D applies to As and Bs. It also depends on features of D's evidential situation that are beyond their control. In particular, if D's evidence makes it easier for them to identify qualified As than qualified Bs—if there are more "clear" cases among the As and more "marginal" cases among the Bs[31]—then we would expect D's success rate to be higher for As than Bs, even if D uses the same decision threshold for As and Bs and otherwise proceeds in an entirely unbiased way. As Simoiu et al. (2017) explain,

> Outcome tests … are imperfect barometers of bias. To see this, suppose that there are two, easily distinguishable types of white drivers: those who have a 1% chance of carrying contraband, and those who have a 75% chance. Similarly, assume that black drivers have either a 1% or 50% chance of carrying contraband. If officers, in a race-neutral manner, search individuals who are at least 10% likely to be carrying contraband, then searches of whites will be successful 75% of the time whereas searches of blacks will be successful only 50% of the time. This simple example illustrates a subtle failure of outcome tests known as the problem of *infra-marginality* ….[32]

Equalized Odds is a close cousin of outcome-based tests for taste-based discrimination, but differs from them in two respects. First, it compares *actual* success rates for predictions about whether decision subjects are qualified, rather than *observed* success rates. Second, it is a test for procedural unfairness in general, and not just than taste-based

[30] The claim that fairness requires using the same decision-threshold across social groups is sometimes called the "single-threshold rule." See Mayson (2019), Huq (2019), and Corbett-Davies and Goel (2018) for discussion. See also footnote 22.

[31] As Hedden puts it; see p. 225.

[32] Simoiu et al. (2017), 1994.

discrimination in particular. So Equalized Odds is essentially an idealized, generalized version of an outcome-based test. One might then worry that it inherits the problem of infra-marginality.

Hedden develops a highly abstract case that seems to show that it does. In Hedden's case, we imagine that twenty people occupy two rooms, A and B, and are each carrying biased coins whose weights match their labels. In room A, twelve people carry coins labeled "0.75" and eight people carry coins labeled "0.125." In room B, ten people carry coins labeled "0.6" and ten people carry coins labeled "0.4." Suppose that a binary classifier labels everyone with a coin weighted 0.5 or higher "heads," and everyone with a coin weighted lower than 0.5 "tails." In this case, the false-positive rate for room A is 3/10, while the false-positive rate for room B is 4/10. Similarly, the false-negative rate is 1/10 for room A, but 4/10 for room B. Both conjuncts of Equalized Odds, then, are violated.[33] Despite this, Hedden argues that the classifier is perfectly fair: "there is seemingly no unfairness of any kind anywhere in this situation."[34]

Hedden takes the example to show that an algorithm can violate Equalized Odds without treating anyone unfairly. Further, the argument is not merely supposed to apply to toy cases like the coin case, but is instead is supposed to generalize to show that Equalized Odds is not a necessary condition of fairness in real-world cases with serious moral stakes, such as predicting recidivism risk for purposes of making pretrial detention decisions.

Now—as I mentioned above—Hedden's argument concerns the fairness of predictions rather than decisions, and my version of Equalized Odds concerns the latter rather than the former. However, his argument can be generalized to apply to my version as well. All we need do is suppose that the algorithm in question is being used to solve some qualification problem. Let's suppose, then, that the algorithm in question is being used to estimate recidivism risk for purposes of deciding whether to detain defendants pretrial. As before, we can assume for concreteness that defendants qualify for pretrial detention if and only if their objective chance of recidivism is above 50%. Let's also assume that the judge in question detains a defendant if and only if the algorithm predicts that they will recidivate (which it will do if their risk of recidivism is greater than 50%).

Let's take the room that a defendant is in to represent their race: defendants in room A and B are black and white respectively. What about the weights of their coins? We know that a defendant's coin is supposed to represent the probability that they will recidivate. However, Hedden does not specify whether we should understand these probabilities as

[33] Hedden (2021), 221-222.

[34] Hedden (2021), 220.

objective chances or evidential probabilities. I will consider both interpretations in turn.

Suppose we take the weight of a defendant's coin to represent their objective chance of recidivism. This immediately yields the problem that was raised in section 3 for the problem of ideal accuracy: given what has been said about this case, the decision procedure described will never misclassify anyone as either qualified or unqualified for detention, and so will not violated Equalized Odds.[35] So the objection fails if we take the weights of decision subjects' coins to represent objective chances.

Suppose alternatively that the weight of a defendant's coin represents the probability that they will recidivate conditional on the available evidence. Now we get a different problem, which is that the case is underspecified. To determine whether any given defendant is misclassified as qualified (or unqualified), we need to know that defendant's objective chance of recidivating. Since we know the weight of each person's coin, we know their evidential probability of recidivating. Hedden also has us assume that "actual relative frequencies match coin biases," which means we know that, of the defendants whose coin has a weight of x%, x% of those defendants will actually recidivate. But this is consistent with a wide variety of assumptions about how objective chances of recidivism are distributed across the forty defendants in the case. In particular, it is consistent with the assumption that the algorithm's risk scores match each defendant's objective chance of recidivism—which would mean that, as before, using the algorithm would *not* lead to a violation of Equalized Odds.

To generate a counterexample to Equalized Odds featuring dispositional qualifications, we need a case in which evidential probabilities of recidivism come apart from the associated objective chances in such a way that Equalized Odds is violated despite the decision-maker proceeding in a seemingly unbiased way. The following case, *Defendants*, will do the trick.

Suppose that, as in Jewel Thieves, you are a judge deciding which defendants to detain pretrial on the basis of recidivism risk. However, this time God is not on hand to tell you each defendant's objective chance of recidivism. Instead, the angel Gabriel—who is known to be infallible but not omniscient—appears to you and offers to assist you in deciding which defendants to detain. After studying recidivism extensively, Gabriel has determined that defendants are always either high risk (with a 90% objective chance of recidivism) or low risk (10%). Gabriel does not have an

[35] What if we chose an example featuring categorical qualifications instead of dispositional ones? Now we cannot construct a case that is structurally analogous to Hedden's, because the relevant objective chances will all be 0% or 100%. (And even if we could, the algorithm would still not violate Equalized Odds, because it would still classify everyone correctly.)

infallible way to determine which are which, but has developed an algorithm that classifies defendants into one of two categories, A or B. He stresses that whether a defendant belongs to category A or B is not determined by reference to their race or clearly objectionable statistical proxies for race such as residing in a predominantly white neighborhood. Instead, membership in A or B is determined on the basis of risk factors whose causal relationship to recidivism is not mediated by race. Gabriel tells you that, on average, 80% of individuals in category A are objectively high risk, whereas 80% of individuals in category B are objectively low risk. It follows that defendants in category A have an evidential probability of recidivism of 74%; for defendants in category B, it is 26%.

You are responsible for deciding whether to detain each of 200 defendants, 100 black and 100 white. Gabriel tells you that, as it happens, all of the black defendants belong to category A and all of the white defendants belong to category B. (Assume that no other evidence of recidivism risk is available.) If you then go on to judge all and only defendants whose evidential probability of recidivism exceeds 50% to be qualified for detention, then you will judge all black defendants to be qualified and all white defendants to be unqualified. The result will be a dramatic violation of Equalized Odds. On the one hand, the expected false-positive rates (in my sense) for black and white defendants will be 100% and 0%. On the other hand, the expected false-negative rates for black and white defendants will be 0% and 100%. However, it seems clear that you will not be engaging in taste-based discrimination against black defendants: your estimates of recidivism risk will simply be those that are dictated by the evidence available to you, and you will be applying the same decision threshold to both groups. Moreover, the suggestion that you are treating black defendants unfairly seems odd. How could it be unfair to treat each defendant as the available evidence demands they be treated?

Does this show that Equalized Odds is not a requirement of procedural fairness? It does not. Recall our earlier observation that qualification problems are cases of imperfect procedural fairness. In such cases, there is an independent standard for when decisions are *substantively* fair, but no perfectly reliable procedure for producing such decisions is available. Real-world qualification problems are cases of just this kind. A particular decision in a qualification problem is substantively fair iff the person is qualified for the benefit/burden and receives it, or unqualified for the benefit/burden and does not. However, decision-makers never have infallible access to whether particular individuals are qualified. While *perfect* procedural fairness in solving real-world qualification problems is thus out of reach, we can nonetheless evaluate available decision procedures in terms of how closely those procedures *approximate* the ideal of perfect procedural fairness. (As Rawls says, "The fundamental criterion for judging any procedure is the justice of its likely

results."[36]) And the decision procedure featured in Defendants does fall short of the ideal of perfect procedural fairness in a quite dramatic way. Objectively low-risk black defendants *have no chance at all* of being granted the pretrial release they are qualified to receive, whereas objectively high-risk white defendants are *guaranteed* to avoid the pretrial detention they qualify for. This is a significant departure from perfect procedural fairness, which means that Defendants is not a case in which a decision procedure violates Equalized Odds despite being ideally procedurally fair.

I suspect that the temptation to say that your decision procedure in Defendants is fair stems from the thought that if the decision procedure you use was *not* fair, it would also be morally wrong for you to use it. It does seem implausible that you are doing anything wrong by proceeding as you do in Defendants, because the alternatives all seem worse from a moral point of view. For example, you could try to equalize false-positive and false-negative rates by randomly judging some white defendants to be high risk and detaining them, or randomly judging some black defendants to be low risk and releasing them. Or you could set the evidence that Gabriel has provided aside, and simply detain or release all defendants regardless of the evidence. All of these options seem morally worse than using your chosen decision procedure. But if using that procedure is your morally best option, then it is hard to accept that using it would be morally wrong, and so that using it would be unfair, as Equalized Odds would have it.

To answer this objection, we need to clarify what it means to say that Equalized Odds is a requirement of procedural fairness. What I mean is this: any decision procedure that violates Equalized Odds is thereby at least pro tanto procedurally unfair, in the sense that it falls short of the ideal of perfect procedural fairness. It does not follow that using such a procedure would be morally wrong: Equalized Odds does not operate as a "hard constraint" on decision-makers' choice of procedures for solving qualification problems. But Equalized Odds need not be a hard constraint of this kind to have important implications for how decision-makers ought to act. In my view, since Equalized Odds is a requirement of procedural fairness, decision-makers have a pro tanto duty to avoid violating it. If decision-makers violate this duty without adequate moral justification (such as that violating it is the morally best option overall), then their behavior will be all-things-considered morally wrong, and not merely pro tanto unfair.

Moreover, Equalized Odds has implications for how decision-makers ought to act even in cases where the morally best decision procedure currently available violates Equalized Odds. In these cases, decision-makers have a pro tanto reason of procedural fairness to seek out

[36] Rawls (1999), 202.

methods for assessing decision-subjects' qualifications that will enable them to better satisfy Equalized Odds. For example, in Defendants you have a pro tanto reason of procedural fairness to ask Gabriel if there is any way for you to do a better job distinguishing high- and low-risk black defendants. In real-world decision-making contexts, knowing that the morally best decision procedure currently available violates Equalized Odds gives decision-makers a pro tanto reason of procedural fairness to invest in the development of improved methods for assessing qualifications that will ameliorate the problem. If decision-makers fail to make such investments without an adequate excuse, then it seems reasonable to say that they are thereby wronging those who are adversely affected by the disparity, in virtue of failing to take appropriate precautions to ensure that they are treated fairly.

What are the implications of the foregoing discussion for Hedden's objection? The motivating thought behind Hedden's objection seems to be this: due to the phenomenon of infra-marginality, it is possible that following the available evidence where it leads—by treating each decision-subject as the evidential probabilities dictate—will nonetheless result in a violation of Equalized Odds. This initially seems to show that Equalized Odds cannot be a requirement of procedural fairness. If you simply treat each decision-subject in the way that the available evidence suggests they ought to be treated, how can anyone complain that they are being wronged? The answer to this challenge is that it can be pro tanto unfair to use a decision procedure even if doing so wrongs no one. In cases of infra-marginality of the kind that are supposed to generate a problem for Equalized Odds, deficiencies in the available evidence make it impossible for decision-makers to achieve perfect procedural fairness. They are thus *not* cases in which even a perfectly fair decision procedure would violate Equalized Odds, and so are not counterexamples to the claim that Equalized Odds is a requirement of procedural fairness.

6 Conclusion

In this paper, I have developed a new version of Equalized Odds and shown that it avoids two key problems for the criterion, the problem of ideal accuracy and the problem of infra-marginality. I will conclude by emphasizing two more general morals of the preceding discussion, briefly addressing the conflict between Equalized Odds and Calibration Within Groups, and identifying a few questions that require further research.

Two more general lessons of my argument bear emphasis. First, the statistical criteria of fairness that computer scientists have recently proposed are supposed to help us determine whether the predictive algorithms that institutions use to allocate important burdens and benefits treat decision subjects fairly. Many of the proposed criteria place constraints on how particular kinds of predictive errors (such as false

positives) are distributed across different kinds of individuals (such as black and white defendants). One upshot of my defense of Equalized Odds is that what counts as a predictive mistake from the perspective of procedural justice depends on the normative structure of the decision problem the predictive algorithm in question is being used to solve. Unless we attend carefully to that normative structure, we are likely to apply statistical criteria of fairness incorrectly—as ProPublica did—with misleading results.

Second, the foregoing discussion shows that procedural unfairness can occur as a result of deficiencies in the evidence available to a decision-maker, as opposed to deficiencies in how the decision-maker makes decisions on the basis of that evidence. I propose that we call this species of procedural unfairness *evidentiary unfairness*. It seems to me that evidentiary unfairness is likely to be widespread, and likely to have an outsized impact on members of marginalized social groups—making it an important obstacle to social justice.[37]

Chief among the questions about Equalized Odds that the foregoing discussion does not address is what we should make of the conflict between Equalized Odds and Calibration Within Groups, another popular statistical criterion of fairness. The conflict between the two criteria has received considerable attention in the literature, and is sometimes taken to give us a reason to reject Equalized Odds.[38] While I

[37] Why would evidentiary unfairness disproportionately affect members of socially marginalized groups? Because social marginalization tends to generate evidence that affected individuals lack the traits that qualify them for more favorable treatment by social institutions. Crucially, this evidence is generated even in cases where the individuals in question *do* qualify for favorable treatment. For example, some widely accepted risk factors for recidivism are in effect measures of an individual's "level of legitimate economic opportunity" (to borrow a poignant phrase from COMPAS' user manual). Using these features to estimate recidivism risk will presumably lead courts to judge low-risk individuals from marginalized groups to be higher-risk than their counterparts from more privileged backgrounds. Similar points apply, mutatis mutandis, to other features institutions treat as a basis for allocating benefits and burdens. (Consider taking the prestige of job applicants' undergraduate institution into account in making hiring decisions, or taking wealth into account in making lending decisions.)

[38] Equalized Odds and Calibration Within Groups are normally incompatible when base rates of the feature of interest differ across groups. This has been demonstrated formally for the version of Equalized Odds discussed by ProPublica; see Kleinberg et al. (2016), Chouldechova (2017), and Miconi (2017). The conflict also arises for my revised version of Equalized Odds as it applies to both dispositional and categorical qualification problems. (To see that the problem arises in dispositional cases, consider that racial profiling will be necessary to ensure Calibration Within Groups is satisfied in cases where race is an independent risk factor, in the sense that it gives us evidence of objective risk that

lack the space to give the issue adequate treatment here, the preceding discussion suggests a way forward.

I have argued that taking Equalized Odds to constitute a pro tanto requirement of fairness (rather than a hard constraint) helps us to see that the problem of infra-marginality does not pose a serious challenge to the criterion. It seems plausible that Calibration Within Groups is also a pro tanto requirement, if it is a requirement of fairness at all.[39] The need to adjudicate conflicting pro tanto duties of procedural fairness is familiar from the legal epistemology literature. For example, one issue in the design of fair procedures for conducting criminal trials is the conflict between the duty to find guilty defendants guilty and the duty to find innocent defendants innocent; the famous "Blackstone ratio" provides guidance about how to balance these two competing demands of procedural fairness. The whole point of the concept of a pro tanto duty is to allow for such conflicts, and it is no argument against the existence of one pro tanto duty that there is another that sometimes (or even normally) conflicts with it. So I do not think that the conflict between Equalized Odds and Calibration Within Groups is a serious problem for my defense of Equalized Odds. That said, the question of whether Calibration Within Groups is a requirement of procedural fairness—and (if so) how it interacts with Equalized Odds—requires further attention.

A few other outstanding issues bear mentioning as well. First, Equalized Odds identifies two distinct constraints on decision procedures used to solve qualification problems that may come into conflict. How should we adjudicate these conflicts, when they arise?[40] Second, which groups count for purposes of applying Equalized Odds? For example, should black women be considered a separate group from black men for purposes of applying the principle? Third, what factors determine the risk thresholds that determine how decision subjects ought to be treated in dispositional qualification problems? Finally, is Equalized Odds grounded in more fundamental principles of procedural justice, and if so, what are they?[41] I hope to explore these questions in future research.

---

is not screened off by other available evidence.) See Long (2021) for an argument that we should accept Calibration Within Groups and reject Equalized Odds.

[39] See Castro (2022) and Eva (2022) for arguments that Calibration Within Groups is not a requirement of fairness.

[40] Long (2021) cites this as a reason to reject Equal False Positive Rates—one half of Equalized Odds—as a requirement of procedural fairness. Long and others raise additional objections to Equalized Odds as well (see footnote 4 for references); I think these objections can be answered, but answering them is beyond the scope of this paper.

[41] See Castro (2019) for one proposal about the normative foundations of Equalized Odds.

Acknowledgements

For helpful feedback on earlier versions of this paper, I would like to thank Arden Ali, Jennifer Carr, Ryan Doody, Jay Hodges, Lily Hu, Gregory Keenan, Euan MacDonald, Greg Ray, Friederike Schuur, Nikita Shepard, and anonymous reviewers at the ACM Conference on Fairness, Accountability, and Transparency. I would also like to thank audiences at several presentations of this paper in 2020 and 2021, including at the International Society for Justice Research Annual Meeting, the University of Edinburgh Legal Theory Research Group Seminar Series, the University of Florida, the University of Florida South Eastern Graduate Philosophy Conference, the Harvard University Cyberethics Forum, the Jain Family Institute, and the Rocky Mountain Ethics Congress. Special thanks to Jenna Donohue, Milo Phillips-Brown, Duncan Purves, and two anonymous reviewers at *Synthese* for extensive feedback on previous drafts.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in- criminal-sentencing.

Ayres (2002). Outcome tests of racial disparities in police practices. *Justice Research and Policy* 4(1–2), 131-142.

Basu, R. (2019a). Radical moral encroachment: The moral stakes of racist beliefs. *Philosophical Issues* 2019(1), 9–23.

Basu, R. (2019b). The wrongs of racist beliefs. *Philosophical Studies* 176(9), 2497–515

Becker, G. S. (1957). *The Economics of Discrimination*. Univ. Chicago Press, Chicago, IL.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *J. Polit. Econ*. 101**,** 385–409.

Castro, C. (2019). What's wrong with machine bias? *Ergo, an Open Access Journal of Philosophy*, *6*.

Castro, C. (2022). Just machines. *Public Affairs Quarterly*, 36 (2), 163–183.

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., Pierson, E., Feller, A., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post*.

Corbett-Davies, S. and Goel, S. (2018), The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint*. *arXiv:1808.00023*.

Di Bello, M., & O'Neil, C. (2020). Profile evidence, fairness, and the risks of mistaken convictions. *Ethics*, *130*(2), 147-178.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS Risk Scales: Accuracy Equity and Predictive Parity. Retrieved from https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Eva, B. (2022). Algorithmic fairness and base rate tracking. *Philosophy & Public Affairs*, *50*(2), 239-266.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Federal Probation*, *80*, 38.

Gardiner, G. (2019). The reasonable and the relevant: Legal standards of proof. *Philosophy & Public Affairs*, *47*(3), 288-318.

Glynn, L. (2010). Deterministic chance. *The British Journal for the Philosophy of Science*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, *49*(2).

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*(4), 811-866.

Huq, A.Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal* 68(6), 1043-1134.

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

List, C., & Pivato, M. (2015). Emergent chance. *Philosophical Review*, *124*(1), 119-152.

Long, R. (2021). Fairness in machine learning: against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy*, *19*(1), 49-78.

Mayson, S. (2019). Bias in, bias out. *Yale Law Journal* 128(8), 2122–2473.

Miconi, T. (2017). The impossibility of "fairness": a generalized impossibility result for decisions. *arXiv preprint arXiv:1707.01195*.

Moss, S. (2018). Moral encroachment. *Proceedings of the Aristotelian Society* 118(2), 177–205.

Rawls, J. (1999). *A Theory of Justice* (revised edition). Belknap Press.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11(3), 1193–1216.

Smith, H. M. (2014). The subjective moral duty to inform oneself before acting. *Ethics*, *125*(1), 11-38.

Strevens, M. (1999). Objective Probability as a Guide to the World. *Philosophical Studies*, 243-275.

Sutton, W., & Linn, E. (2004). *Where the money was: The memoirs of a bank robber*. Crown.