Forthcoming in *Analysis*

An Impossibility Theorem for Base Rate Tracking and Equalised Odds

Rush Stewart Benjamin Eva Shanna Slank Reuben Stern

February 12, 2024

Abstract

There is a theorem that shows that it is impossible for an algorithm to jointly satisfy the statistical fairness criteria of Calibration and Equalised Odds non-trivially. But what about the recently advocated alternative to Calibration, Base Rate Tracking? Here, we show that Base Rate Tracking is strictly weaker than Calibration, and then take up the question of whether it is possible to jointly satisfy Base Rate Tracking and Equalised Odds in non-trivial scenarios. We show that it is not, thereby establishing an even more general impossibility theorem.

Keywords. AI ethics; algorithmic fairness; bias; base rate tracking; calibration; equalised odds

1 Introduction

What does it mean for a predictive algorithm to treat groups fairly? One influential approach to answering this question involves identifying *statistical criteria of algorithmic fairness*, i.e., necessary conditions that must be satisfied by the statistical profile of an algorithm's predictions in order for the algorithm to count as 'fair.' Three prominent alleged examples of such criteria are Calibration, Equalised Odds, and Base Rate Tracking. Kleinberg et al. (2017) famously prove that Calibration is incompatible with Equalised Odds in all realistic cases. We show below that Calibration implies Base Rate Tracking,[1] and then take up the question of whether Base Rate Tracking is likewise incompatible with Equalised Odds. Our answer takes the form of a general impossibility result regarding the mutual satisfiability of Base Rate Tracking and Equalised Odds. We take this result to show that Base Rate Tracking and Equalised Odds cannot both be necessary conditions for the fairness of a predictive algorithm. Since our result weakens the assumptions of the result due to Kleinberg, et al., ours is a generalisation of their result.

[1] This may seem at odds with Eva's (2022) finding that Base Rate Tracking and Calibration are logically independent, but Eva works with a notion of Calibration that is strictly weaker than the notion that Kleinberg et al. (2017) deploy—namely, what he calls 'Weak Calibration' and Stewart and Nielsen (2020) call 'Predictive Equity.' He does not consider the logical relationship between Calibration as we define it and Base Rate Tracking. See p. 248 of Eva's paper for his reasons for focusing on the weaker notion.

2 The Framework and the Criteria

We begin by introducing the basic formal framework in which the fairness criteria will be articulated.[2] First, we assume that there exists a finite population N of individuals, each of which either have some relevant property y or not. We model this with the random variable $Y : N \to \{0,1\}$ such that $Y(i) = 1$ if i has the property, and $Y(i) = 0$ otherwise. An *assessor* is a function $h : N \to [0,1]$ assigning numbers to individuals. We interpret $h(i)$ as the probability that an algorithm assigns to individual i having the property y. We call $h(i)$ the 'risk score' assigned to individual i by h.

To facilitate talking about proportions or frequencies, it will be helpful to define a uniform probability distribution P over N. For instance, the quantity $P(Y = 1) = \mu$ represents the proportion of people in N that have property y.[3] We call μ the *base rate* for y in N. Any population N can be divided into different groups. For example, we could divide a population into distinct racial groups, or sex groups, or age groups. We will work with *partitions* of N into groups so that each individual belongs to exactly one group. Given a partition $\pi = \{G_1, \dots, G_m\}$ of N, $P_k = P(\cdot|G_k)$ is the uniform probability distribution on G_k for $k = 1, \dots, m$. The quantity $P_k(Y = 1) = \mu_k$, for example, is the base rate for y in group k.

We let $E_G(h)$ denote the expected value of h for members of G. The expectation E_G is defined in terms of the probability distribution P_G. Because P_G is uniform, $E_G(h)$ represents the average value of h for members of G, i.e., the average risk score for G. Finally, we define the generalised false positive rate $f_G^+(h) = E_G(h|Y = 0)$ and the generalised false negative rate $f_G^-(h) = E_G(1 - h|Y = 1)$. The quantity $f_G^+(h)$ is the average risk score among the members of G that do *not* have property y. Similarly, $f_G^-(h)$ is the average of the quantity 1 minus the risk score among the members of G that *do* have the property y.

With this framework in hand, we can introduce the three candidate statistical criteria of algorithmic fairness mentioned before. To illustrate the conceptual motivation for condition, we consider a special case in which π is a partition of the population into racial groups, and G and G' represent the Asian and Hispanic subpopulations, respectively. First, we have Base Rate Tracking.

> **Base Rate Tracking**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $E_G(h) - \mu_G = E_{G'}(h) - \mu_{G'}$.

In the context of the special setting mentioned just above, suppose that while the base rates for y are almost identical for both Asian and Hispanic populations, the average risk score that h assigns to Asians is much higher. This would be a clear violation of Base Rate Tracking, which requires that the deviation of the average risk score from the base rate in a group is the same for all groups.[4] Equivalently, the difference between the average risk scores assigned to

[2] We take the framework directly from Stewart and Nielsen (2020); Stewart (2022). For simplicity of exposition, we restrict explication to the problem of predicting recidivism in the criminal justice system, but the framework and criteria are much more broadly applicable.

[3] We write $Y = 1$ in the expression $P(Y = 1)$ as shorthand for the event $\{i \in N : Y(i) = 1\}$, where events are subsets of N.

[4] It is important to stress that the deviation of the average risk score from the base rate has to be in the *same direction* for all groups—the signs match on the left- and right-hand sides of the central equation in the definition of Base Rate Tracking.

those two groups should be equal to the difference between the base rates of those groups. Essentially, the thought behind Base Rate Tracking is that an assessor is fair only if the average assessment within any group is driven by prevalence of the property within that group *to the same extent that it is for all groups.* To put it another way, groups should only be treated differently to the extent that they are relevantly different, where what counts as relevantly different is captured by the difference in the base rates.

The next criterion is Calibration.

> **Calibration**. For an assessor h of N and any group G in the relevant partition π of N, $P_G(Y=1|h=p)=p$ for all $p \in [0,1]$ such that $P_G(h=p)>0$.

Calibration requires that for any possible risk score p, the percentage of Hispanic subjects who are assigned risk score p who actually have the relevant property should be equal to p. And the same goes for the percentage of Asian subjects with risk score p who actually have the relevant property. So defined, Calibration implies a weaker property sometimes called *Predictive Equity*: the percentage of Hispanic subjects assigned risk score p who have property y is the same as the percentage of Asian subjects assigned risk score p who have property y. Thus, if 20% of Hispanic subjects assigned a risk score of 0.8 in fact have property y but 90% of Asian subjects assigned a risk score of 0.8 have property y, then the assessor violates Predictive Equity. Roughly, the idea is that any given risk score should have the same evidential import for each relevant group, e.g., a risk score of 0.8 should mean the same for Asians as it does for Hispanics. If Asians with a risk score of 0.8 have the relevant property at a much lower rate than Hispanics with a risk score of 0.8, then the risk score 0.8 seems to mean different things for those two groups, and that seems problematic.

To Predictive Equity's requirement that these percentages must be equal across groups, Calibration adds that these percentages must be equal to the value p. For example, suppose that 20% of Hispanics assigned a risk score of 0.8 and 20% of Asians assigned a risk score of 0.8 are recidivists. While this does not violate Predictive Equity, it does violate Calibration. Again, Calibration requires not only that the percentage of recidivists assigned a score of 0.8 is the same for both groups, but that, in either group, 80% of those assigned a score of 0.8 are recidivists. On the one hand, one might view Predictive Equity as concerned exclusively with *equal treatment* of groups, whereas Calibration adds a sort of accuracy requirement (Stewart and Nielsen, 2020, p. 740). We return briefly to this view at the close of the essay. On the other hand, one might view Calibration's additional requirement as demanding an additional sort of fairness. Stewart (2022) discusses a motivation for Calibration in terms of prohibiting forms of over- and under-confidence in the riskiness of different groups. In the example under discussion, the assessor is uniformly over-confident in both Asian and Hispanic recidivism, which is arguably unfair.

Finally, we have Equalised Odds.[5]

[5] Unfortunately, as with Calibration, multiple criteria go by the name 'Equalised Odds' in the literature, leading to many instances of misreported and misattributed results. We note that what we call Equalised Odds is the condition that appears in the result due to Kleinberg et al. (2017). In this context, equal generalised false positive rates is called 'Balance for the Negative Class' and equal generalised false negative rates is called 'Balance for the Positive Class'. The conjunction of these two conditions is dubbed 'Equalised Odds' by Kleinberg and coauthors in Pleiss et al. (2017). We also note that this version of Equalised Odds is weaker than the other commonly discussed criterion that shares its name. Briefly, the stronger criterion requires that, for the negative (resp. positive) class in each group, equal proportions of individuals are assigned any given

Table 1: Calibrated Risk Assessment

Asian	$h(1^*) = 2/3$	$h(2) = 2/3$	$h(3^*) = 2/3$
Hispanic	$h(4) = 1/3$	$h(5) = 1/3$	$h(6^*) = 1/3$

> **Equalised Odds**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $f^+_G(h) = f^+_{G'}(h)$ and $f^-_G(h) = f^-_{G'}(h)$ (whenever those terms are defined).

Equalised Odds requires that the generalised false negative rates and generalised false positive rates should be the same for the Asian subpopulation and the Hispanic subpopulation. That is, the average risk score for Hispanics who *lack* property y is the same as the average risk score for Asians who *lack* property y, and the average risk score for Hispanics who *have* property y is the same as the average risk score for Asians that *have* property y. In the case at hand, Equalised Odds prohibits cases in which, say, Asian subjects who do not reoffend are routinely given higher risk scores than their Hispanic, non-recidivist counterparts. In other words, it prohibits the algorithm from making more favourable mistakes for one group than it does for the other.

To get a feel for how these criteria relate to one another, consider the example illustrated in Table 1, in which there is a population of 6 individuals, partitioned into two racial groups (Asian and Hispanic). Individuals that have the property y are indicated with an asterisk. The base rate for the Asian group is $\frac{2}{3}$, while the base rate for the Hispanic group is $\frac{1}{3}$. The average risk scores assigned to the groups align perfectly with the corresponding base rates: all Asians are assigned a risk score of $\frac{2}{3}$ and all Hispanics are assigned a risk score of $\frac{1}{3}$. In this case, Calibration is perfectly satisfied, but Equalised Odds is violated because (i) the generalized false positive rate for Asians is $f^+_{Asian}(h) = E_{Asian}(h|Y = 0) = \frac{2}{3}$ while the generalized false positive rate for Hispanics is $\frac{1}{3}$, and (ii) the generalized false negative rate for Asians is $f^-_{Asian}(h) = E_{Asian}(1 - h|Y = 1) = \frac{1}{3}$ while the generalized false negative rate for Hispanics is $\frac{2}{3}$. In light of the results of Kleinberg et al. (2017), it is unsurprising that Equalised Odds is violated in this case. We know that Equalised Odds and Calibration can only be jointly satisfied when either (i) the base rates of the relevant groups are equal, or (ii) h is perfect.[6] Since neither of those conditions are satisfied and Calibration is, Equalised Odds cannot possibly be satisfied. Any advocate of the necessity of Calibration must make peace with the ubiquitous violation of Equalised Odds, and vice-versa.

Note that Base Rate Tracking is also satisfied in Table 1. This illustrates that Base Rate Tracking is compatible with Calibration in non-trivial cases. Table 2 illustrates a slight variation on the case in Table 1, wherein Base Rate Tracking is satisfied but Calibration and Equalised Odds are both violated. So what is the precise logical relationship between Base

score p. The weaker criterion just requires that, for the negative (resp. positive) class in each group, the *average* score is the same.

[6] By 'perfect,' we mean that h assigns a risk score of 1 to all agents with the property y and 0 to all agents that lack the property y: $h(i) = Y(i)$ for all $i \in N$.

Table 2: Base Rate Tracking Risk Assessment

Asian	$h(1^*) = 5/6$	$h(2) = 5/6$	$h(3^*) = 5/6$
Hispanic	$h(4) = 1/2$	$h(5) = 1/2$	$h(6^*) = 1/2$

Rate Tracking and Calibration? We can show that Calibration *implies* Base Rate Tracking but, by Table 2, not the converse.

Proposition 1. *If an assessor h for a population N satisfies Calibration for a partition π of N, then h satisfies Base Rate Tracking for π. The converse implication does not obtain.*

In light of Kleinberg et al.'s impossibility theorem, we have good reason to seek ways to weaken Calibration or Equalised Odds if we think that both conditions are getting at important fairness properties that our algorithms ought to satisfy. Proposition 1 says that Base Rate Tracking weakens Calibration. So are there any non-trivial cases in which Base Rate Tracking and Equalised Odds are both satisfied? Or must the advocate of Equalised Odds eschew not only Calibration but also its weaker counterpart, Base Rate Tracking?

3 The Result

We can now state the central result and the main contribution of this paper.

Theorem. *If an assessor h for a population N satisfies Base Rate Tracking and Equalised Odds for some partition π of N, then either h is perfect or the base rates for all groups in π are identical.*

This result is analogous to the results for Calibration obtained by Kleinberg et al. (2017). Actually, given the restriction to population partitions, the theorem generalises Kleinberg et al.'s result by relaxing Calibration to Base Rate Tracking. It establishes that it is impossible to jointly satisfy Base Rate Tracking and Equalised Odds in all non-trivial cases, that is, in all cases in which the base rates of the relevant groups are unequal and the algorithm is not perfect. So the advocate of Equalised Odds has to surrender both Base Rate Tracking and Calibration.

Eva indicates one possible escape route worth considering. Base Rate Tracking is defined in terms of the difference of certain quantities. But we could just as well consider alternative functional forms. For instance, as Eva writes, "one might plausibly reformulate Base Rate Tracking in terms of the ratios of average risk scores and base rates, rather than the differences. The resultant formulation is clearly and importantly distinct from the [difference formulation], although it has the same motivation and is equally able to diagnose the intrinsic unfairness of the predictions [in the specific examples under consideration]" (2022, p. 260, slight formatting revisions for consistency). So we might consider Base Rate Tracking in ratio form.

Ratio Base Rate Tracking. For an assessor h of N and any groups G, G' in the relevant partition π of N, $\frac{E_G(h)}{\mu_G} = \frac{E_{G'}(h)}{\mu_{G'}}$ (whenever both ratios are defined).

Ratio Base Rate Tracking also follows from Calibration (this is immediate from equation 1 in the proof of Proposition 1). But note that while the assessor in Table 2 satisfies Base Rate Tracking, it does *not* satisfy Ratio Base Rate Tracking. So Eva is right that the ratio form of the condition is importantly distinct from the difference form.[7] Is it possible to satisfy *this* condition and Equalised Odds outside of trivial scenarios? No.

Proposition 2. *If an assessor h for a population N satisfies Ratio Base Rate Tracking and Equalised Odds for some partition π of N, then either h is perfect or the base rates for all groups in π are identical.*[8]

So the escape route is blocked, and the impossibility result is robust to the choice of form of Base Rate Tracking.[9]

Like Predictive Equity, (Ratio) Base Rate Tracking relaxes some of the stringency of Calibration while retaining an aspect of Calibration explicitly concerned with equal treatment of relevant groups. To the extent that Calibration requires something over and above a form of fair treatment—such as a sort of accuracy, as one of the interpretations we mentioned in introducing Calibration has it—one might worry that the property is not a compelling candidate criterion of *fairness*. This issue is substantial and requires more attention than we can give it here. But it might suggest that an impossibility result that is framed in terms of Predictive Equity or (Ratio) Base Rate Tracking rather than Calibration, like our Theorem and Proposition 2, is even more straightforwardly relevant to the issue of the joint satisfaction of *fairness* criteria than the original Kleinberg et al. result. So in addition to strengthening that result, our results plausibly bear more directly on the opening question about the meaning of 'algorithmic fairness.'

Acknowledgements: Thanks to Jonathan Herington, Michael Nielsen, Walter Sinnott-Armstrong, two anonymous referees, and an audience at the annual KCL Cumberland Lodge philosophy retreat for helpful discussions and feedback. Rush gratefully acknowledges funding from the Digital Futures Institute at King's.

[7] For an investigation of what differences these different formulations of Base Rate Tracking make to the analysis of actual COMPAS data, see (Crespo et al., MS).

[8] Since the proof of Proposition 2 is fairly similar to the proof of the Theorem, we omit it.

[9] As a referee points out, an alternative strategy for attempting to skirt the sort of impossibilities we report is explored by Reich and Vijaykumar (2020). The key idea is to impose criteria, not jointly on the assessor, but in different ways. As they write about Calibration and a form of Equalised Odds for binary classification, "We relax the mathematical tension between these two fairness criteria by separately enforcing *calibration on the score* and *equal error rates on the corresponding classifier*" (2020, p. 4:2). Since we've shown that Base Rate Tracking is strictly weaker than Calibration, a similar *possibility* result for Base Rate Tracking would follow from Reich and Vijaykumar's observation.

Appendix

Proof of Proposition 1

Proof. Suppose that h satisfies Calibration for π. For any group $G \in \pi$, let

$$h[G] = \{p : h(i) = p \text{ for some } i \in G\}$$

be the image of G under h. Using the law of total probability and Calibration,

$$\begin{aligned} \mu_G &= P_G(Y = 1) \\ &= \sum_{p \in h[G]} P_G(Y = 1 | h = p) P_G(h = p) \\ &= \sum_{p \in h[G]} p P_G(h = p) \\ &= E_G(h) \end{aligned} \tag{1}$$

From 1, it follows that, for any $G \in \pi$, $E_G - \mu_G = 0$. Hence, for any $G, G' \in \pi$, $E_G(h) - \mu_G = E_{G'}(h) - \mu_{G'}$. So Calibration implies Base Rate Tracking.

Table 2 is an example in which h satisfies Base Rate Tracking but not Calibration, which establishes that Base Rate Tracking does not imply Calibration. □

Proof of the Theorem

Proof. Suppose that h satisfies Base Rate Tracking and Equalised Odds for a partition π of N. For *reductio*, suppose that h is not perfect and that not all base rates are identical. Let $\mu_G, \mu_{G'}$ witness this latter assumption. By the law of total expectation, and using the linearity of expectation,

$$\begin{aligned} E_G(h) - \mu_G &= (E_G(h|Y = 0)(1 - \mu_G) + E_G(h|Y = 1)\mu_G) - \mu_G \\ &= E_G(h|Y = 0) - E_G(h|Y = 0)\mu_G + E_G(h|Y = 1)\mu_G - \mu_G \\ &= E_G(h|Y = 0) - \mu_G(E_G(h|Y = 0) + 1 - E_G(h|Y = 1)) \\ &= E_G(h|Y = 0) - \mu_G(E_G(h|Y = 0) + E_G(1 - h|Y = 1)) \\ &= f^+_G(h) - \mu_G(f^+_G(h) + f^-_G(h)). \end{aligned} \tag{2}$$

And, similarly,

$$E_{G'}(h) - \mu_{G'} = f^+_{G'}(h) - \mu_{G'}(f^+_{G'}(h) + f^-_{G'}(h)). \tag{3}$$

By Base Rate Tracking, 2 and the right hand side of 3 are equal:

$$f^+_G(h) - \mu_G(f^+_G(h) + f^-_G(h)) = f^+_{G'}(h) - \mu_{G'}(f^+_{G'}(h) + f^-_{G'}(h)). \tag{4}$$

By Equalised Odds, we can rewrite the right hand side of 4 as

$$f^+_G(h) - \mu_G(f^+_G(h) + f^-_G(h)) = f^+_G(h) - \mu_{G'}(f^+_G(h) + f^-_G(h)),$$

which implies

$$\mu_G(f_G^+(h) + f_G^-(h)) = \mu_{G'}(f_G^+(h) + f_G^-(h)). \tag{5}$$

Since we have assumed for *reductio* that h is not perfect, for some G^*, at least one of $f_{G^*}^+(h)$ and $f_{G^*}^-(h)$ is positive which implies that the sum $f_{G^*}^+(h) + f_{G^*}^-(h)$ is positive. By Equalised Odds, the same is true for every group in π. So, the sum in 5 is positive. It follows that $\mu_G = \mu_{G'}$, which contradicts our assumption that μ_G and $\mu_{G'}$ witness that not all base rates are equal. Thus, either h is perfect or all base rates are equal. □

Remark 1. Perfection implies both Equalised Odds and Base Rate Tracking. Suppose that h is perfect. It follows that all error rates for groups in π are 0 (and hence 5 does not imply equal base rates in this case), so h satisfies Equalised Odds. And by 2, $E_G(h) - \mu_G = 0$ for each group in π, so h satisfies Base Rate Tracking.

Remark 2. While the Theorem and Proposition 2 suggest a general result along the lines of Theorem 2 in (Stewart and Nielsen, 2020), our results are not corollaries of that particular result. Neither Base Rate Tracking nor Ratio Base Rate Tracking implies property $*$.

References

Crespo, V., B. Eva, and W. Sinnott-Armstrong (MS). Applying base rate tracking and COMPAS. Unpublished Manuscript.

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

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.

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.

Reich, C. L. and S. Vijaykumar (2020). A possibility in algorithmic fairness: Can calibration and equal error rates be reconciled? *arXiv preprint arXiv:2002.07676*.

Stewart, R. T. (2022). Identity and the limits of fair assessment. *Journal of Theoretical Politics 34*(3), 415–442.

Stewart, R. T. and M. Nielsen (2020). On the possibility of testimonial justice. *Australasian Journal of Philosophy 98*(4), 732–746.