Forthcoming in *Philosophy & Technology*

Spanning in and Spacing out? A Reply to Eva

Michael Nielsen[1] and Rush T. Stewart[2]

[1]The University of Sydney
[2]University of Rochester

November 20, 2024

In our "New Possibilities for Fair Algorithms," the key to avoiding the famous impossibility result for Calibration and Equalized Odds (Kleinberg et al., 2017) is to replace Calibration with a weaker condition we call *Spanning*. Spanning requires that, for each relevant group, an assessor's predictions capture the group base rate in the sense that the base rate lies within the interval spanned by the assessor's forecasts. We are grateful for Benjamin Eva's critical and constructive engagement with our proposal.

Eva is responsible for what has so far been the most interesting fairness criterion proposed in the philosophy literature: Base Rate Tracking (Eva, 2022). In his comment on our paper, he emphasizes the "intra-group" nature of Spanning—it imposes a constraint on the assessments within each group rather than requiring some parity in assessment to hold across groups—and suggests an alternative to Spanning that he dubs *Spacing*. Spacing is essentially a form of intra-group Base Rate Tracking.

> **Spacing**. For any relevant protected group G, the difference between G's base rate and the average risk score assigned to G should be no greater than some threshold $t \in [0, 1]$.

Eva points out that Spacing is similar to Spanning in that "(i) it encodes a similar ideal, (ii) can be applied to individual groups in isolation," but differs insofar as it "(iii) is seemingly able to diagnose some apparent cases of unfairness that Spanning does not identify" (slight formatting alteration for consistency). Eva then puts to us the key question: "One salient question then is whether there is any reason to focus on Spanning rather than Spacing." We would like to make three points in reply.

Our first point is that the question might rest on a false dilemma. There is an important sense in which we do not have to choose between Spanning and Spacing. Even in the presence of Equalized Odds, Spanning and Spacing are consistent. To see this, observe that any assessor satisfies Spacing if and only if $t \geq \max_{G \in \pi} |E_G(h) - \mu_G|$. So the same goes for assessors that satisfy Equalized Odds and Spanning, of which we know there are many (Nielsen and Stewart, 2024, Theorem 1). The threshold t need only fall in the appropriate range. This raises a crucial question for defenders of Spacing. How *should* the threshold be selected? Eva concedes that that this is an open issue for Spacing and pivotal for its plausibility. We note one salient fact that we return to below: when $t = 0$, Spacing implies Base Rate Tracking which, unlike Spanning, is inconsistent with Equalized Odds outside of trivial cases (Stewart et al., 2024, Theorem).

Second, we would like to remark on the dialectical import of point (iii) in the quotation above. Eva considers an example that we discuss (p. 11) of apparently biased assessment that nevertheless satisfies Spanning. He observes that, at least for certain choices of t, Spacing is violated by the example, so there is a form of bias that Spacing diagnoses and Spanning does not. But, for one thing, the diagnosis depends crucially on the value of t, which we haven't been told how to choose yet. Moreover, there are other forms of bias that Spanning diagnoses and Spacing does not. For instance, for any value of $t \in (0, 1]$, the assessor that is constant within each group G, with $h(i) = \min(\mu_G + t, 1)$ for every $i \in G$, satisfies Spacing. Yet these predictions exhibit the sort of uniform overconfidence in each group that motivated our introduction of Spanning. So Eva's point (iii) does not seem to us to indicate any comparative advantage over Spanning after all. Finally, and most importantly, all of this is as we would expect. As we say originally about the example under discussion, we are happy to concede that assessors satisfying Spanning can be biased because Spanning is only plausible as a *necessary* condition of fair assessment, not as a sufficient one. No more could be claimed with any plausibility—and Eva does not claim otherwise—for Spacing. Counterexamples of this sort can really be put only to full accounts of fair assessment.

Third, and finally, if one insists on viewing Spanning and Spacing as competitors, there are considerations—even if not dispositive—that favor Spanning. One of the three motivations we give for Spanning is the attractiveness of its ideal. A criterion's *ideal* is a condition equivalent to that criterion's satisfaction for *all* partitions of the population; put differently, a criterion's ideal characterizes maximal fairness according to that criterion. Spanning shares the ideal of perfect assessment with Calibration and Eva's own Base Rate Tracking. On some ways of developing the "ideals program" in algorithmic fairness (Stewart, 2024), this is significant content from Calibration that Spanning retains. Perfect assessment is plausibly a sufficient condition for fair assessment (Hardt et al., 2016; Stewart and Nielsen, MS). One idea sympathetically considered in (Stewart, 2024) is that having an ideal that is sufficient for fairness is a necessary condition for an overall account of algorithmic fairness. Putting these two ideas together, any account that includes Calibration, Base Rate Tracking, or Spanning will not face ideal-based objections since an ideal sufficient for fairness will be implied by any such account. But criteria with ideals that clearly leave room for unfairness assume an additional burden since defending ideal-based objections requires specifying additional criteria of fairness. The addition of any of a range of standard criteria considered in the literature will not suffice since they are either logically weaker than perfect assessment in ways that leave room for unfairness (Stewart, 2024, Figure 1) or are ethically objectionable (Stewart and Nielsen, MS, Problem 3). Eva claims that Spacing's ideal is similar to Spanning's, and it is. But it is not identical. In fact, it's strictly weaker. Say that an assessor is t-Perfect when $t \geq |h(i) - Y(i)|$ for all $i \in N$.

Observation. *An assessor h for a population N satisfies Spacing with respect to t for all partitions iff h is t-Perfect.*

For positive values of t, t-Perfection clearly does leave room for systematic bias. For instance, a recidivism assessor that assigns every black non-recidivist the score t and all black recidivists a score of 1 while assigning all white non-recidivists 0 and all white recidivists the score $1 - t$ is t-Perfect but exhibits systematic bias. How dramatic the failure of fair assessment is depends on the value of t. When $t = 0$, on the other hand, Spacing implies Base Rate Tracking which has been shown to be consistent with Equalized Odds only when assessment is perfect or all group base rates are the same (Stewart et al., 2024, Theorem). Whatever the value of t, then, Spacing lacks important properties that Spanning has: an ideal sufficient for fairness or non-trivial consistency with Equalized Odds.

References

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

Hardt, M., E. Price, and N. Srebro (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems 29*.

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.

Nielsen, M. and R. T. Stewart (2024). New possibilities for algorithmic fairness. *Philosophy & Technology 37*(116).

Stewart, R., B. Eva, S. Slank, and R. Stern (2024). An impossibility theorem for base rate tracking and equalised odds. *Analysis*, Forthcoming.

Stewart, R. T. (2024). The ideals program in algorithmic fairness. *AI & Society*, Forthcoming.

Stewart, R. T. and M. Nielsen (MS). What's wrong with statistical parity? Unpublished Manuscript.