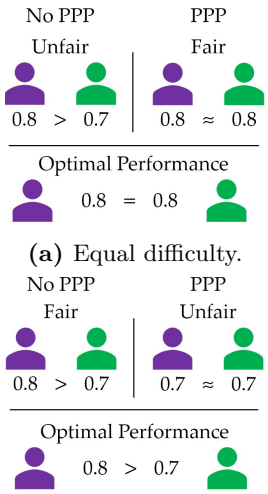
Parity in predictive performance is neither necessary nor sufficient for fairness

Justin Engelmann, Miguel O. Bernabeu*, Amos Storkey*

TL;DR: Parity in Predictive Performance (PPP) holds that a machine learning model is fair if and only if its predictive performance (by some measure) is (approximately) equal across groups of interest [1]. We argue that this assumes that groups are equally difficult, which is unlikely to hold in practice. Absent this assumption, a model could be fair but not satisfy PPP, or be unfair yet satisfy PPP. Thus, PPP is neither necessary nor sufficient for fairness. We propose a new definition of fairness, Relative Realised PPP (R2P3), to account for these situations.

Intro: PPP is a class of fairness definitions that look at the parity of some measure of predictive performance across groups [1]. A number of PPP definitions have been proposed that simply differ in the measure they use [2]. PPP is especially sensible in settings such as medical diagnosis where high predictive performance is in the interest of the decision subject. It is easy to evaluate and is intuitively appealing: If our model performs much worse on a group of people, then it appears that we are failing them. However, this implicitly assumes that we could be doing better for this group, which in turn - we argue - stems from the assumption that groups are equally hard. Thus, higher performance for one group would imply that the same performance is achievable for all groups, and not doing so would be unfair. This case is illustrated in Figure (a). Intuitively, the left model is unfair and the right model is fair; and PPP is indeed necessary and sufficient for this sense of fairness.



(a) Equal difficulty.

(b) Different difficulty.

Differences in group difficulty: However, groups might not be equally hard. For example, in a medical setting, a protected attribute (e.g. sex, race) might cause anatomical differences that make diagnosis harder in one group than in another. In other settings such as criminal justice, we only observe some of the variables that cause the outcome. If the relative importance between the observed and unobserved variables for the outcome differs across groups, this too will make some groups easier than others.

Failure of PPP: This case is illustrated in Figure (b). Here, we would consider the left model fair as it achieves optimal performance for both groups, and the right model unfair to the purple group. However, the left model does not satisfy PPP, so PPP is not necessary for fairness. The right model, on the other hand, does satisfy PPP but is unfair, thus PPP is not sufficient for fairness. Thus, differences in group difficulty break PPP.

Relative Realised PPP (R2P3): We propose R2P3 as a definition of fairness that accounts for this. R2P3 holds that a model is fair if and only if there is (approximate) parity in terms of how much of the optimal performance has been realised across groups. This can account for our moral intuitions in both of the cases that we presented. In fact, R2P3 is a generalisation of PPP and recovers PPP for the special case where groups are equally difficult. R2P3 is more complex and nuanced than PPP, but fairness itself is a complex, nuanced concept. In our opinion, after reflecting on the possibility of differences in group difficulty and that this appears likely in many practical contexts, R2P3 is both intuitive and has desirable implications: R2P3 requires that groups are treated equally in terms of effort expended towards performing well for the given group, which then constitutes fair treatment.

Implications & future work: The notion of an "optimal performance" is - in our opinion - sensible, but might be hard to quantify in practice. We could develop a model per group that aims to maximise performance in that group only. Doing this every time a model is developed would be wasteful in terms of expended resources. Instead, this could be done thoroughly for specific problems (e.g. skin cancer diagnosis) once to then inform our idea of the difference in group difficulty each time a model is developed for the same problem (or a closely related one). We think this is a promising direction for future work. The argument presented here also implies that this work is necessary: We should not naively compare group averages of performance to judge model fairness, unless we have strong reason to believe that there are no differences in group difficulty.

[1] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[2] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.