Parity in Predictive Performance is neither necessary nor sufficient for fairness

Justin Engelmann[12]✉, Miguel O. Bernabeu[1]∞, Amos Storkey[2]∞

✉ justin.engelmann@ed.ac.uk ∞Equal supervision

[1] Centre for Medical Informatics, Usher Institute, University of Edinburgh [2] Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh

Background: Parity in Predictive Performance (PPP)

- Parity in Predictive Performance (PPP) across groups of interest is a commonly used definition of algorithmic fairness.
- Moral intuition: If our algorithm performs worse for one group than others, then that's unfair to that group.
- Example: Suppose an AI-based medical decision support system was less accurate for one sex/ethnicity/age group than others. This seems obviously unfair and undesirable.

Diagnostic AI model: +

Two groups of interest:

Performance per group: 0.8 > 0.7

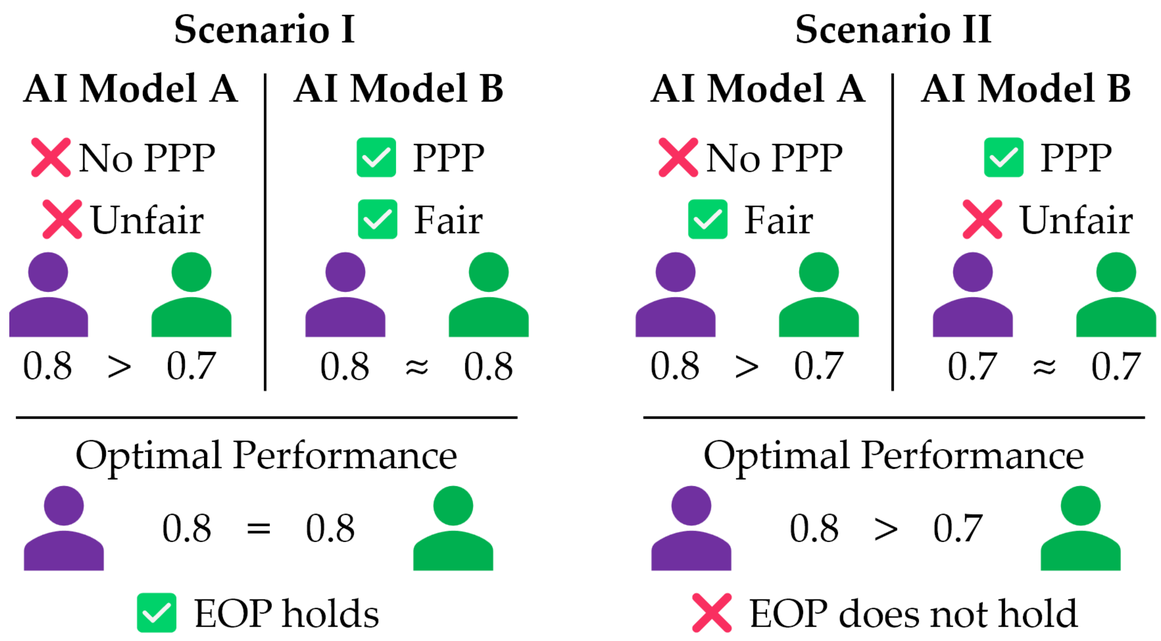
The implicit assumption of PPP: Equal Optimal Performance (EOP)

✗ This doesn't seem fair!

- PPP implicitly assumes that all groups are equally difficult, i.e. the optimal performance is the same across groups.
- With EOP, the better performance in group is evidence that we could do better for group, too.
- This appears to be the reason why the scenario above does not seem fair. We're failing the group!

Parity in Predictive Performance is neither necessary nor sufficient for fairness

- When the EOP holds, PPP does indeed seem to be both necessary and sufficient for fairness. This is illustrated in Scenario I, where EOP holds and in our moral intuition Model A is unfair and Model B is fair.
- But – as illustrated in Scenario II –when EOP does not hold, then we can construct scenarios where a model does not satisfy PPP but is fair, or satisfies PPP but is not fair.
- Hence, Parity in Predictive Performance is neither necessary nor sufficient for fairness!



Is Equal Optimal Performance (EOP) a reasonable assumption?

- PPP is only a good definition of fairness if EOP holds, as seems to be commonly implicitly assumed. But does this assumption stand up to scrutiny? We argue that it does not and often groups are unlikely to be equally difficult.
- Why might the Optimal Performance (OP) differ across groups?
  - A) Group attributes are **essentially** linked to OP.
    Example: In medicine, protected attributes like age, sex, or ethnicity can cause physical differences between groups. Those in turn can affect how fundamentally easy or difficult diagnosis of a specific disease is. Thus, OP might differ.
  - B) Group attributes are **circumstantially** linked to OP.
    Example: A protected attribute like ethnicity can correlate with other variables (e.g. age due to migration patterns, social deprivation/drug abuse due to systemic racism), and these other variables can in turn be essentially linked to OP (e.g. predicting disease risk in young people is typically a very different problem from doing so in older people; cardiac arrests due to drug abuse are different from those that aren't).

Relative Realised Parity in Predictive Performance (R2P3)

- R2P3 is a generalisation of PPP where the parity of interest is not in absolute predictive performance but in relative realise predictive performance, i.e. group absolute performance relative to group OP.
- R2P3 can thus account for our moral intuitions both in the case that EOP holds and in the case that it does not.
- Objection: Group OP is typically an unobservable quantity, making R2P3 more challenging to apply than PPP.
- Reply: True, but A) this does not change the fact that PPP falls short of being a good definition and B) we can try to reason about expected group OPs domain knowledge or try to estimate them for a class of problems by – for example – trying to build one AI model per group, each with the goal of maximising performance of the given group.

Conclusions & Future work

- Despite its popularity, PPP assumes EOP which is often doubtful and at least non-obvious in many cases.
- R2P3 overcomes these difficulties but is harder to apply as it needs knowledge of group OP.
- Future work should further explore R2P3 conceptually and develop methods/frameworks for approximating group OP in practice.

This work was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics.