Classification Parity, Causal Equal Protection and Algorithmic Fairness

Marcello Di Bello[1], Nicolò Cangiotti[2] and Michele Loi[2]

[1] *Arizona State University, 975 S. Myrtle Ave P.O.Box 874302, Tempe, AZ 85287, United States*
[2] *Polytechnic University Milan, via Bonardi 9, Campus Leonardo, 20133 Milan, Italy*

Keywords

Algorithmic Fairness; Classification parity; Equal protection; Counterfactual Fairness; Statistical Profile; Diagnostic Evidence

1. Introduction

The literature in computer science and philosophy has formulated several criteria of algorithmic fairness. One of these is *classification parity* also known under different names. It requires (roughly) that people who belong to different socially salient groups (say groups defined by race or gender) should have the same prospects of erroneous positive and negative classification by a predictive algorithm. At first blush, this is an intuitive criterion of algorithmic fairness. A number of authors in the legal and philosophical literature, however, have argued that classification parity is not a criterion of algorithmic fairness we should take seriously [1, 3, 5]. On the other hand, independently of discussions about algorithmic fairness, other authors [2] have defended an analogous yet different principle—equal protection—that applies to decisions in criminal trials. This principle requires that the risks of a mistaken positive classification (say, a conviction) be equal across factually negative (say, innocent) individuals who belong to different relevant groups. Equal protection is a form of classification parity for false positives in which the true value of the target variable is "innocence" and the "relevant group" is picked out by any feature used as a statistical profile.

Given the similarity between classification parity and equal protection, we explore the relationships between the two. In particular, we seek to address two questions. First, is equal protection threatened by the criticisms of classification parity as a plausible criterion of algorithmic fairness? Conversely, if equal protection can be defended as a criterion of algorithmic fairness, to what extent does this contribute to make classification parity plausible *in general* as a principle of fairness for prediction-based decisions? To keep the discussion manageable, we focus on classification parity in the context of trial decisions to which equal protection was originally intended to apply.

2. Causal Equal Protection

Inspired by discussions in Kusner [4], Hedden [3] Long [5], and Beigang [1] we begin by drawing a distinction between classification parity and what we will call *causal* equal protection. Unlike classification parity, causal equal protection requires that innocent individuals not be exposed to higher risks of conviction *because* they belong to a specific profiled group. The two requirements are not the same: suppose that judges' decisions are in part guided by the statistical profile that *young* individuals are more likely to commit criminal acts. In this case, classification parity can be violated across two

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland

EMAIL: mdibello@asu.edu (A. 1); nicolo.cangiotti@polimi.it (A. 2); michele.loi@polimi.it (A. 3)
ORCID: 0000-0001-8684-8827 (A. 1); 0000-0003-2200-446X (A. 2); 0000-0002-7053-4724 (A. 3)

CEUR Workshop Proceedings (CEUR-WS.org)

groups, say Orange and Green, even if this violation only happens because Orange includes a higher proportion of young individuals compared to Green. Instead, causal equal protection would not be violated under similar circumstances so long as Orange individuals are not more likely to be misclassified *because* they are Orange. But, causal equal protection would be violated if the feature of being Orange were to guide trial decisions. It would also be violated even when the feature of being Orange was not intentionally used as a criterion in trial decisions, but decisions to convict were not counterfactually fair [4] for members of Orange and Green. In other words, causal equal protection can be viewed as counterfactual fairness conditional on the target variable.

3. Statistical Profiles

With the distinction between classification parity and causal equal protection in hand, we ask whether the intuitive unfairness of violating classification parity or causal equal protection depends on whether a statistical profile guides the decisions as opposed to a decision that does not rely on statistical profiles.

We begin by drawing a distinction between statistical profiles (or statistical evidence) and diagnostic evidence based on causal ordering. Predictive evidence lies upstream in the causal structure relative to the target variable, while diagnostic evidence lies downstream. When diagnostic evidence is used to guide a decision, the casual path from group membership to the decision is typically mediated by the target variable. When predictive evidence is used, the causal path from group membership to the decision is unmediated by the target variable. So the use of a group characteristic (e.g., Orange) as a statistical profile will violate causal equal protection. But no such violation should occur when diagnostic evidence guides the decision: insofar as the evidence is causally downstream relative to the target variable, any causal influence of the group membership onto the decision would be blocked by conditioning on the target variable.

We argue that cases in which statistical profiles are deployed in decision-making are clearly unfair because*,* by construction, they violate causal equal protection (and not merely classification parity).

4. A Pro Tanto Principle

We conclude by showing that causal equal protection can be defended from the objections in Hedden [3] and Long [5] against classification parity. When classification parity is violated *because statistical profiles are used* (as opposed to diagnostic evidence), causal equal protection is violated and this violation is *pro tanto* unfair irrespective of the nature of the group in question. In contrast, when classification parity is violated by a decision-making process that relies on diagnostic information, it is implausible to regard this violation as morally problematic in the general case. The pro tanto reasons for blocking predictions causally influenced by group features must, however, be weighed against consequentialist considerations, including the loss of predictive accuracy and their distributive effects.

5. References

[1] F. Beigang (2023), Reconciling Algorithmic Fairness Criteria, *Philosophy and Public Affairs* 51

[2] M. Di Bello & C. O'Neil (2020), Profile Evidence, Fairness, and the Risks of Mistaken Convictions, *Ethics* 103(2).

[3] B. Hedden (2021), On Statistical Criteria of Algorithmic Fairness, *Philosophy & Public Affair* 49(2).

[4] M. Kusner et al. (2018), Counterfactual Fairness, preprint: https://arxiv.org/abs/1703.06856.

[5] R. Long (2021), Fairness in Machine Learning: Against False Positive Rate Equality, *Journal of Moral Philosophy* 19(1).