A Human-in-the-Loop Fairness-Aware Model Selection Framework for Complex Fairness Objective Landscapes

Jake Robertson[1, 2], **Thorsten Schmidt**[1], **Frank Hutter**[1, 3], **Noor Awad**[1]
[1]University of Freiburg,
[2]Zuse School ELIZA,
[3]ELLIS Institute Tübingen
robertsj@cs.uni-freiburg.de, thorsten.schmidt@stochastik.uni-freiburg.de,
fh@cs.uni-freiburg.de, awad@cs.uni-freiburg.de

Abstract

Fairness-aware Machine Learning (FairML) applications are often characterized by complex social objectives and legal requirements, frequently involving multiple, potentially conflicting notions of fairness. Despite the well-known Impossibility Theorem of Fairness and extensive theoretical research on the statistical and socio-technical trade-offs between fairness metrics, many FairML tools still optimize or constrain for a single fairness objective. However, this one-sided optimization can inadvertently lead to violations of other relevant notions of fairness. In this socio-technical and empirical study, we frame fairness as a many-objective (MaO) problem by treating fairness metrics as *conflicting objectives*. We introduce *ManyFairHPO*, a human-in-the-loop, fairness-aware model selection framework that enables practitioners to effectively navigate complex and nuanced fairness objective landscapes. ManyFairHPO aids in the identification, evaluation, and balancing of fairness metric conflicts and their related social consequences, leading to more informed and socially responsible model-selection decisions. Through a comprehensive empirical evaluation and a case study on the Law School Admissions problem, we demonstrate the effectiveness of ManyFairHPO in balancing multiple fairness objectives, mitigating risks such as self-fulfilling prophecies, and providing interpretable insights to guide stakeholders in making fairness-aware modeling decisions.

1 Introduction

Instances of algorithmic discrimination are a growing concern in both the machine learning (ML) literature and, more recently, in the media and greater society. This is a consequence of the increasing prevalence of ML applications where individuals are disparately impacted by algorithmic decisions (Angwin et al. 2016; de Zwart 2022). Mirroring the complex and socio-technical nature of the machine bias problem, the field of Fairness-aware Machine Learning (FairML) has emerged, providing a collaborative space for political philosophers, social scientists, legislators, statisticians, and ML researchers. FairML has the overarching goals of defining, studying, detecting, and mitigating algorithmic bias.

Despite significant advancements, the FairML community has received widespread criticism from the Social Sciences, Humanities, and Law for attempting to solve the complex, nuanced, and socio-technical problem of machine bias *algorithmically* (Hoffmann 2019; Selbst et al. 2019). Several arguments cite that real-world applications of FairML are often characterized by a complex set of social objectives and legal requirements (Ruf and Detyniecki 2021). Due to their complexity, these criteria are unlikely to be captured by single, coarsely-grained statistical measures of fairness. In such cases, FairML methods that only incorporate a single notion of fairness risk *fair-washing*, or proposing a so-called *fair* model that satisfies one notion of fairness while violates another potentially relevant one, potentially resulting in negative social consequences. Other criticisms of FairML cite the black-box nature of bias-mitigation techniques as a key concern (Robertson, Stinson, and Hu 2022), suggesting a crucial interplay between fairness, transparency, and interpretability (Barocas, Hardt, and Narayanan 2023; Schöffer 2023). Broadly effective FairML methods should not only cope with a diverse set of objectives and requirements but ideally offer interpretable insights.

Rather than resisting these criticisms, we embrace them, taking the perspective that FairML approaches that oversimplify the complex and socio-technical nature of FairML problems risk doing more harm than good. Instead, FairML approaches must adapt to the context in which they exist. In recent years, the topic of fairness has gained popularity in the Automated Machine Learning (AutoML) literature, which typically formulates fairness as a bi-objective (BiO) or constrained hyperparameter optimization (HPO) problem. Fairness-aware HPO varies common ML design decisions (tree depth, neural architecture, neural network width, etc.) to explore the Pareto Front of fair and accurate models (Wu and Wang 2021; Perrone et al. 2021; Schmucker et al. 2020; Dooley* et al. 2023). According to Weerts et al. (2023), fairness-aware AutoML holds key advantages in human-centricity and transparency, enabling practitioners to explore multiple fairness-accuracy trade-offs and gain interpretable insights into the fairness-objective landscape of the problem at hand. In addition, the BiO problem formulation extends naturally to the MaO case, encompassing three or more fairness and accuracy constraints or objectives, enabling practitioners to explore not only fairness-accuracy trade-offs but trade-offs between fairness metrics themselves. Although the prospect of constraining or op-