On Group and Individual Fairness

***Abstract*—This paper discusses the impact of our chosen paper on the current state of bias in machine learning as well as the present and future developments of fair machine learning.**

I. A Discussion on Our Chosen Paper

A. Group and Individual Fairness Definitions

Our chosen paper [1] address two fundamental fairness metrics used in machine learning (ML) to measure a model's bias. Group fairness (GF) compares protected groups whilst individual fairness (IF) compares similar cases. GF metrics are widely used whilst IF metrics are uncommon as they require a mathematical definition of *task-specific similarity* which is harder to formalise. Previously, it was thought that ensuring GF contradicts IF and vice versa [2], our paper addresses this misconception.

B. Group and Individual Fairness Are Not in Conflict

The paper proposes a new insight which follows from the definition of *luck egalitarianism* – that justice should be determined by choices made of someone's own free will (*effort*), and not by the differences in their circumstance (*luck*). If we can identify the unjust logical assumptions made on our model, we can represent the same bias with both GF and IF if we carefully choose which metrics to implement whilst respecting our assumptions.

For example, in one scenario provided by our paper, it is suggested to construct an IF metric that adjusts the values of features within groups that are affected by a group disparity due to structural discrimination.

Some obvious theoretical complications occur such as identifying the assumptions made on the model. Similarly, implementational discrepancies occur such as – regarding the example metric – quantifying the size of the adjustment and on which features to even apply this correction. All these intricacies are highly subjective and currently require human decision makers which are still inherently biased.

C. How We View Group and Individual Fairness

Our paper summarises the differences as being between *worldviews* rather than between GF and IF. Two opposing worldviews suggested are *we're all equal* (WAE) and *what you see is what you get* (WYSIWYG) [3].

Understanding that these contradictions occur between specific metrics rather than between GF and IF themselves, will allow for a better consideration of which metrics to implement when evaluating a model's bias. Similarly, we will be able to development more sophisticated algorithms for mitigating both GF and IF simultaneously, due to a greater understanding of where the true conflicts arise in these definitions. It may also help identify currently used biased models that were previously overlooked; comparable with real cases [4] which have only recently been highlighted as biased.

However, the restrictions described in the previous section prevent the paper immediately having a direct impact on bias mitigation without some implementational formalisation that may require relaxation or reconsideration of the given notions. For example, it has already been proposed that neither of the two worldviews provided necessarily hold in practice which motivated the suggesting of a third worldview *α-Hybrid* [5].

II. The Present and Future State of Fair Machine Learning

A. The Present State of Fair Machine Learning

Most current bias mitigation algorithms apply to the former of the two types of ML – *supervised* and *unsupervised* learning. Identifying and mitigating fairness for unsupervised learning is an important area of research [6] as the extent of the influence of unsupervised learning is still widely unknown, and as such, may have a strong influence on human-centric problems.

Within many current GF mitigation algorithms, we discover inconsistencies in the fairness of overlapping protected groups – this is known as *fairness gerrymandering*. One of the main problems with mitigation algorithms that consider combinations of multiple groups to reduce fairness gerrymandering, is that they become increasingly complex, do not scale well, and suffer from overfitting as the number of groups increases [7]. This requires further consideration, despite lying within the 'simplistic' notion of GF.

B. The Future of Fair Machine Learning

Our concept of IF originates from Aristotle's maxim stating that alike cases should be treated alike. However, a related maxim he also proposed on *individual justice* is in direct contrast with our definitions of GF and even IF despite the similar intuitions [8]. It states that all cases need to be assessed individually without using historical correlations [9]. In fact, this definition conflicts with ML itself which relies on training data to make predictions. If we can formalise this definition such that it is appropriate for ML, it may become a useful metric for determining bias.

Humans have many innate *cognitive biases* of which some are capable of having a positive influence on ML [10]. However, it is worth considering the possibility of adopting some of the more problematic cognitive biases – specifically in the design of *artificial general intelligence* which is modelled on the human brain.

Amongst the discussion on the fair part of ML, it is easy to lose sight of the goal of ML itself: to create a model that accurately predicts on a given task. Many bias mitigation algorithms lead to a trade-off between fairness and bias. In a fair machine learning utopia, we would expect to see a perfectly fair model with no cost to accuracy. This is not necessarily realistic – especially given the philosophical discussions regarding the definition of 'fairness' – but always needs to be considered when approaching bias in ML.

[1] R. Binns, “On the apparent conflict between indiviual and group fairness”

[2] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, “Fairness through awareness"

[3] S. A. Friedler, C. Schneidegger, S. Venkatasubramanian, “On the (im)possibility of fairness”

[4] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”

[5] S. Yeom, M. C. Tschantz, “Avoiding disparity amplification under different worldviews”

[6] F. Chierichetti, R. Kumar, S. Lattanzi, S. Vassilvitskii, “Fair clustering through fairlets”

[7] M. Kearns, S. Neel, A. Roth, Z. Wu, “Preventing fairness gerrymandering: Auditing and learning for subgroup fairness”

[8] R. Binns, “Human judgment in algorithmic loops: Individual justice and automated decision-making”

[9] F. Schauer, “On treating unlike cases alike”

[10] H. Taniguchi, H. Sato, T. Shirakawa, “A machine learning model with human cognitive biases capable of learning from small and biased datasets”