From Implicit to Explicit Assumptions: Why There is No Fairness Without Bias-Awareness.

MARCO FAVIER, University of Antwerp, Belgium
TOON CALDERS, University of Antwerp, Belgium

This extended abstract is a follow-up to our previous work –*"Patriarchy Hurts Men Too." Does Your Model Agree? A Discussion on Fairness Assumptions.* [2]– We discuss why implicit assumptions for fairness are tied to specific properties of the bias present in the data and why, without explicit assumptions, the choice of the correct model might be difficult. Moreover, we state a new result on one of these possible assumptions, proving the validity of the approach.

Keywords: Algorithmic Bias, Assumption Modeling, Classification

Reference Format:
Marco Favier and Toon Calders. 2025. From Implicit to Explicit Assumptions: Why There is No Fairness Without Bias-Awareness.. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 4 pages.

1 Introduction

The field of algorithmic fairness is, at its heart, a field based on ethical assumptions. For instance, when data show an unexpected disparity between different communities that is not in line with our ethical beliefs, we *assume* that the data are biased or at least reflect the bias in our society. These assumptions become more explicit when we choose a specific fairness metric, as fairness metrics are often based on distinct ethical principles.

Take *Demographic Parity* [1], for example: mathematically speaking, it simply requires the model output ($\widehat{Y}$) to be independent of the sensitive attribute (A). As a mathematical expression, it can be written as $P(\widehat{Y} \mid A) = P(\widehat{Y})$, but ethically, we are making several assumptions. We assume that the communities represented by the sensitive attribute are equally able to perform the task expressed by the output label, that all communities should have equal access to the benefit provided by the model, and that the model's output should not contain information that could be used for demographic profiling.

The use of the word *"assumption"* in this discussion is not meant to be pejorative: we do not use it to mean "something we believe and yet it is false" but rather as a technical term to indicate statements that we do not expect, a priori, to be investigable using the available data. It is not "data does not show it" but rather "data *cannot* show it." For instance, when we use statistical parity as a fairness metric, if the model we trained does not satisfy $P(\widehat{Y} \mid A) = P(\widehat{Y})$, we do not question whether different communities deserve unequal treatment, nor do we abandon our ethical principles. Instead, we question the fairness of the model and explore how to align it with our principles.

Authors' Contact Information: Marco Favier, University of Antwerp, Antwerp, Belgium, marco.favier@uantwerpen.be; Toon Calders, University of Antwerp, Antwerp, Belgium.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

The choice of a fairness metric comes with a set of assumptions that are rather *explicit*: each fairness metric has specific objectives and enforces clear ethical principles that are sufficiently clear to practitioners. However, the choice of a model and how to address bias in the data is often based on more *implicit* assumptions. Just as explicit assumptions can be expressed mathematically through fairness metrics, implicit assumptions, when formalized, can define mathematical properties of the data and the bias present in it.

2 "Patriarchy Hurts Men Too." Does Your Model Agree? A Discussion on Fairness Assumptions.

In our previous work [2], we discuss some implicit fairness assumptions and their consequences. To do so, we consider two probability distributions on the set of individuals $x \in X$:

- $p_f(y_i \mid x)$, which is the probability that an individual $x \in X$ is assigned label y_i from a set of possible labels $Y = \{y_1, \dots, y_n\}$. This is the fair distribution, unaffected by bias, but we are unable to observe it directly.
- $p_u(y_i \mid x)$, which is a probability distribution over the same variables but different from the previous one. This distribution is considered unfair because bias has affected it. It can be directly inferred from the available data.

Having two distinct distributions allows us to evaluate any model according to each probability. Given two models $\widehat{Y}, \widehat{Y}'$, we write $\widehat{Y} \leq_f \widehat{Y}'$ if $P_f(\widehat{Y} = y_i \mid Y = y_i) \leq P_f(\widehat{Y}' = y_i \mid Y = y_i)$ for all labels y_i. Similarly, we write $\widehat{Y} \leq_u \widehat{Y}'$ when using the unfair distribution. Essentially, we compare the models' accuracy when evaluated on the fair and unfair distributions. In particular, a model is called unbiased if it is maximal according to the order on the fair distribution.

Using this framework, we were able to define the following set of implicit assumptions, which we report here, limited to the binary ($Y = \{0,1\}$) classification case for convenience:

Definition 2.1. Let $\widehat{Y}, \widehat{Y}'$ be two models on X, and $A \subseteq X$ any sensitive group of the population X. We refer to the following properties by the names given below:

- **No Harm, No Foul**: $\widehat{Y}|_A <_u \widehat{Y}'|_A \Rightarrow \widehat{Y}|_A <_f \widehat{Y}'|_A$.
- **Fairness as Optimization**: $\widehat{Y}|_A <_f \widehat{Y}'|_A \Rightarrow \widehat{Y}|_A <_u \widehat{Y}'|_A$.
- **Representation Matters**: If $\widehat{Y}\big|_A$ is not a maximal model according to $\leq_u$, then $\widehat{Y}$ is biased.
- **Affirmative Action**: If $\widehat{Y}$ is an unbiased model, there exists a threshold t such that $\widehat{Y}\big|_A = \mathbb{1}_{[t,1]}(p_u(y = 1 \mid x))$.
- **Double Standard**: If $\widehat{Y}$ is a maximal model for $\leq_u$, there exists t such that $\widehat{Y}\big|_A = \mathbb{1}_{[t,1]}(p_f(y = 1 \mid x))$.

where $\mathbb{1}_{[t,1]}$ is the indicator function on the interval $[t,1]$.

A more in-depth discussion of these assumptions, their meanings, and their mutual relationships can be found in [2]. Seemingly, these assumptions only describe how we can construct a fair model from the knowledge of the unfair distribution. For instance, the *No Harm, No Foul* assumption states that if we limit our models to a single sensitive group, any improvement according to unfair probabilities should also be an improvement according to fair probabilities, since under this assumption, "fairness" can be defined only by comparing different sensitive groups.

Our main result in [2] was the following connection, which we state here in a simplified form for the binary classification case:

THEOREM 2.2. *Let $A \subseteq X$ be a sensitive group. Under any of the assumptions in Definition 2.1, for any $x, x' \in A$ we have:*

$$p_f(y = 1 \mid x) \leq p_f(y = 1 \mid x') \iff p_u(y = 1 \mid x) \leq p_u(y = 1 \mid x')$$

That is, within each sensitive group, the fair and unfair distributions are in a monotonic relationship with each other.

This is an important result, as it shows that if we make any of the assumptions in Definition 2.1 explicit, we are already making an assumption about the bias present in the data, and as a consequence, about the correct model to use for debiasing the data. On the other hand, some choices of models might already imply some of these assumptions, which means that we are excluding certain possible biases from the start. Specifically, these assumptions exclude the possibility of within-group bias [4], and without making these assumptions explicit, we might end up using a model that is incapable of addressing it.

3 Connecting Assumptions and Biases

Theorem 2.2 is valid for all the previously stated assumptions, yet this does not mean that all the assumptions are equivalent to each other. Some assumptions are stronger than others, which means it may be possible to find stronger results similar to Theorem 2.2 for each assumption.

The aforementioned theorem can be restated by saying that the space of functions that transform fair probabilities into unfair probabilities (a.k.a. the set of all possible biases) is the set of monotonic functions. Under a stronger assumption, the set of possible biases becomes a subset of the previous one. We have a correspondence between assumptions and set of biases:

$$\begin{array}{ccc} \text{Assumption } A & \xleftrightarrow{1:1} & \{\text{Possible biases for Assumption } A\} \\ \Downarrow & & \downarrow \\ \text{Assumption } B & \xleftrightarrow{1:1} & \{\text{Possible biases for Assumption } B\} \end{array}$$

This approach allows us to explore the space of possible biases by examining the space of assumptions, and vice versa. For example, we can ask under what assumptions the set of biases consists solely of linear functions, or, given a specific assumption, what the properties of the possible biases are. This is important because different types of biases, such as label and selection bias, have distinct mathematical properties [3]. Exploring this connection can help us better understand the bias in the data and how to address it effectively.

4 A Proof of Concept: No Harm No Foul

If we are working on a binary classification problem, we can explicitly define the set of possible biases under the *No Harm, No Foul* assumption. In this case, we can prove the following:

THEOREM 4.1. *Let $A \subseteq X$ be a sensitive group. Under the **No Harm, No Foul** assumption, the set*

$$\{\textit{Biases for } \textbf{\textit{No Harm, No Foul}}\} := \{\beta : [0,1] \to [0,1] \textit{ s.t. } \beta(p_f(y = 1 \mid x)) = p_u(y = 1 \mid x) \textit{ for all } x \in A\}$$

is the following:

$$\{\beta \textit{ s.t. there exist } 0 \leq b \leq a \leq 1 \textit{ s.t. } \beta(p_f(y = 1 \mid x)) = a p_f(y = 1 \mid x) + b p_f(y = 0 \mid x) \textit{ almost everywhere}\}$$

SKETCH OF THE PROOF. We already know that any β must be monotonic. To check the other properties, we can first consider the cases where β is linear and notice that the properties are satisfied. If β is not linear, we can construct a counterexample by taking two different linear approximations at different points. □

5 When Assumptions Can Be Tested.

As mentioned in the introduction, fairness assumptions are statements that we do not expect to be directly reflected in the data. The reason for this is that we are aware of the biases present in the data, which can interfere with it. Thus, the presence of bias justifies the inability to validate the assumptions. However, when the assumptions pertain to the bias itself, we should be able to test them. The logic behind this reasonable expectation is almost tautological: bias is defined as the reason why we cannot validate our assumption. If we construct a model of bias based on certain assumptions, but our data contradict the bias model, there are two possibilities. Either our model is incorrect, or, by the definition of bias, it indicates the presence of additional bias that affects our ability to validate the model. In this case, our model must have been incomplete, as there is still some bias we were not accounting for.

Hence, we should be able to test our assumptions about the bias present in the data. This is not an easy task, as we only have access to the unfair distribution, but there are statements that we can test. For instance, this has been done in [3]. Similarly, we can apply this to Theorem 4.1: as a consequence of its statement, unless bias is the identity function (which means there is no bias), the interval $[a,b]$ of values that $p_u(y = 1 \mid x)$ can take is strictly contained within the interval of values that $p_f(y = 1 \mid x)$ can take. In particular, $[a,b]$ is a strict subset of $[0,1]$. This is a statement that depends only on the data and can be tested.

6 Conclusions

In this short paper, we discussed the importance of making explicit the assumptions about the bias present in the data. We showed how formalizing these assumptions allows us to explore the space of possible biases and how to address them. In particular, we argue that assumptions about bias are the only fairness assumptions that we can realistically expect to test. This makes the study of bias extremely important, and results like Theorem 2.2 presented in [2] and Theorem 4.1 presented in this paper are crucial for understanding the bias present in the data and how to address it. We believe a thorough study of the connection between assumptions and biases is necessary to make progress in the field of algorithmic fairness, and we believe these discussions can find a place in the EWAF community.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[2] Marco Favier and Toon Calders. 2024. " Patriarchy Hurts Men Too." Does Your Model Agree? A Discussion on Fairness Assumptions. *BIAS 2024 4th Workshop on Bias and Fairness in AI* (2024).

[3] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? a study of label and selection bias. *Machine Learning* 112, 12 (2023), 5081–5104.

[4] Sofie Goethals, Marco Favier, and Toon Calders. 2024. Reranking individuals: The effect of fair classification within-groups. arXiv:2401.13391 [cs.LG] https://arxiv.org/abs/2401.13391