On the Moral Justification of Statistical Parity

Corinna Hertweck
corinna.hertweck@zhaw.ch
Zurich University of Applied Sciences,
University of Zurich

Christoph Heitz
christoph.heitz@zhaw.ch
Zurich University of Applied Sciences

Michele Loi
michele.loi@ibme.uzh.ch
University of Zurich

ABSTRACT

A crucial but often neglected aspect of algorithmic fairness is the question of how we justify enforcing a certain fairness metric from a moral perspective. When fairness metrics are defined, they are typically argued for by highlighting their mathematical properties. Rarely are the moral assumptions beneath the metric explained. Our aim in this paper is to consider the moral aspects associated with the statistical fairness criterion of independence (statistical parity). To this end, we consider previous work, which discusses the two worldviews "What You See Is What You Get" (WYSIWYG) and "We're All Equal" (WAE) and by doing so provides some guidance for clarifying the possible assumptions in the design of algorithms. We present an extension of this work, which centers on morality. The most natural moral extension is that independence needs to be fulfilled if and only if differences in predictive features (e.g., ability to perform well on a job, propensity to commit a crime, etc.) between socio-demographic groups are caused by unjust social disparities and measurement errors. Through two counterexamples, we demonstrate that this extension is not universally true. This means that the question of whether independence should be used or not cannot be satisfactorily answered by only considering the justness of differences in the predictive features.

1 INTRODUCTION

A look at current practices suggests that in order to evaluate the "fairness" of a given machine learning model, so-called *fairness metrics* have to be computed. However, this disregards the crucial steps which should precede the calculation of fairness metrics: discussing the moral reasons underlying the decision to select one or more fairness metrics to be enforced. This is easily forgotten as the procedure is not as clear as computing statistical metrics, but much harder. It might require discussions with the stakeholder of an application and finding a compromise – after all, there are very few cases where everyone can agree on the correct choice of fairness metric, in particular because it has been shown that some of them are conflicting [3, 11]. It is thus important to not only provide mathematical definitions of fairness metrics, but to provide some guidance for how to reason about them from a moral perspective.

One popular metric is called *independence*, often referred to as *statistical parity*. While existing literature at first sight often seems to reason about independence in moral terms, a lot of the arguments are either not backed up by moral philosophy or turn out to be purely mathematical. We note that the need for enforcing independence is rarely justified from a philosophical perspective, and that the two spaces (philosophy and mathematics) are often conflated, in part also due to terminology. In this paper, we want to make a contribution towards resolving this ambiguity, and to highlight the relation between mathematical justifications for choosing independence and the corresponding moral significance.

We begin by defining independence mathematically in Section 2 and then reconstruct arguments on when independence is considered the correct fairness metric in Section 3. We will show that these arguments, while at first sight appearing to hold moral value, are actually purely mathematical if taken literally. However, since they suggest that there are moral reasons for choosing independence, we will provide a natural extension for the arguments found in the literature (Section 4). We will then argue that this natural extension is not always in line with our moral intuitions about fairness in specific cases (Section 5). We conclude in Section 6 that the question whether independence should be chosen or not is not sufficiently answered by considering the social injustices occurring between the Potential Space (PS) and Observed Space (OS).

2 WHAT IS INDEPENDENCE?

The arguably most commonly used category of fairness metrics is referred to as *group fairness*, focusing on the question whether socio-demographic groups are treated similarly or receive similar outcomes [7, 15]. Group fairness is tested with respect to specific socio-demographic groups, differentiated through a *sensitive attribute*, which we will denote as A. One of the more prominent fairness metrics falling into the category of group fairness is *statistical parity* [3].

The concept is easiest to explain for binary classification ($\hat{Y} \in 0, 1$) and two groups $A = a$ and $A \neq a$: In such a case, statistical parity requires that the probability of the predicted outcome being positive, i.e., $\hat{Y} = 1$, is equal for $A = a$ and $A \neq a$. In other words, the selection rate $P(\hat{Y} = 1)$ has to be independent of the value of the sensitive attribute. This can be expressed as $P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A \neq a)$ [27]. This formula can be generalized for the case of not only a binary predictor, but any predictor R and possibly more values for the sensitive attribute A: $R \perp A$ [3, 17]. This general proposition is referred to as the fairness criterion *independence* [3]. Other prominent group fairness measures are *Separation* and *sufficiency* [3].

Independence has been particularly influential in the early stages of algorithmic fairness [4, 20] while the two other fairness criteria have become more popular in recent work. The debate around separation and sufficiency was sparked by Angwin et al.'s investigative article *Machine Bias* [1], which gained lots of public attention. Subsequent publications have pointed out what is now known as the *impossibility theorem*: Except for in highly constrained cases, we can only satisfy one of the three fairness criteria [3, 11]. This impossibility theorem forces us to pick a specific fairness criterion to enforce or to find a trade-off between them, which raises the question of when one criteria should be chosen over the other two.

In this paper, we will focus on the moral reasons for implementing independence.

A note about language. When we refer to independence with respect to a predictor, we do not mean to describe a property of this predictor, i.e., we are not stating that it is a fact that a certain decision happens to be statistically independent from a variable. Instead, we are concerned with independence as a normative choice about the way the decision *should* be, something desirable. More precisely, we are not concerned with independence *per se* but with the *prescription* to enforce independence.

If we find that there are moral reasons that require us to implement independence, this means that we have to enforce it. Unsurprisingly, it is rare that we can observe independence in a machine learning model that was trained to optimize e.g., accuracy as independence was not a learning goal. However, we can enforce independence, for example, by making independence a learning constraint. This idea is what we will refer to with the phrase "independence should be used". We can either enforce achieving independence fully or partially, e.g., because we want to trade off fairness with another goal such as business interests, or the utilitarian moral goal to maximize the number of lives saved.

3 WHEN SHOULD INDEPENDENCE BE USED?

The question of when independence should be chosen over separation and sufficiency cannot be answered from a purely mathematical or technical perspective. Friedler et al. [9] propose a framework which enables its users to clarify the *worldview* assumed in the context of their application. Yeom and Tschantz [28] build on Friedler et al. to add rules for when independence should be chosen as a fairness metric instead of separation or sufficiency. This section will discuss both papers and distill their rules for when to apply independence. We also argue that the framework presented by Friedler et al. is insufficient to represent the philosophical debate surrounding independence and thus propose an extension of their framework.

3.1 Existing Framework

The premise of Friedler et al.'s framework is that when a decision has to be made by using data-driven predictions, this prediction is based not on the features that we would ideally have access to, but on proxies. In order to clarify the theoretical explanations, we will work with an exemplary scenario throughout this paper, which to some extent has also been used in [9]: hiring. In this case, the task is to predict employee productivity and take a hiring decision based on this prediction.

The main result of [9] is the distinction between three spaces:

- the Construct Space (CS), which consists of the features that we **want** to base the decision on,
- the OS, which consists of the features that we **actually** base the decision on because the CS is not observable, so the OS is our proxy for the CS, and
- the Decision Space (DS), which encompasses the predictions based on the OS.

Let us now consider these spaces in the context of the hiring example. We borrow the language for this from [18]. Jobs are offered and then filled by *selectors*. Selectors pick whom to hire from the pool of *candidates*, i.e., the people who apply for the job. In order to make this decision, they try to predict who will perform best when hired. Each candidate brings certain *qualifications*, based on which selectors wish to make their choice. However, these qualifications are not directly observable (e.g., how good they are at selling the company's product, how well they fit into the existing team, etc.). Instead, they only have access to noisy representations of these qualifications, which we will refer to as *proxies*. These proxies typically include the CV, the motivation letter, the impressions from the interviews etc. As the selectors only have access to the qualifications through the proxies, they have to base their decision about whom to hire on the proxies. In this example, the qualifications are equivalent to the unobservable CS while the proxies represents the observable OS.

Friedler et al. present two opposing worldviews and advocate for being transparent about which one the prediction model adheres to. The two opposing worldviews are:

- "What You See Is What You Get" (WYSIWYG), which assumes that there are barely any differences between the OS and the CS, meaning that the OS is a good proxy for the CS, implying that observed differences between socio-demographic groups correspond to actual differences. In our example, this means that the usage of CVs, interviews etc. as the proxy for qualifications does not harm or benefit one group more than another.
- Measurement Bias (MB), which assumes that the mapping of individuals from the CS to the OS introduces disparities between the socio-demographic groups, implying that differences between groups in the OS are bigger than in the CS. For the hiring example, this means that using CVs as proxies of qualifications harms one group compared to another one.

Friedler et al. refers to the second worldview as *structural bias*. However, we will call this worldview MB in order to distinguish it from the informal usage of the term "structural bias."

Furthermore, Friedler et al. propose the axiom "We're All Equal" (WAE), which oftentimes aligns with the worldview MB. WAE assumes "that in the construct space all groups look essentially the same" and "that there are no innate differences between groups" [9, p. 8]. If we assume that there are no innate differences in the abilities of socio-demographic groups to perform well on a job, but measure differences once we evaluate their CVs, we will see these observed differences as a result of the MB.

3.2 Our Extension of the Framework

When discussing the cause of group differences in the OS, Friedler et al.'s framework quickly reaches its limits. The main issue is that two different spaces are conflated and merged in the CS. Although Friedler et al. recognize this, they justify not differentiating between the two by saying that a differentiation would still lead to "the same mathematical outcome" [9, p. 8].

However, we believe that it is necessary to clearly distinguish these two spaces as it increases the understanding of the decision making process and helps navigate the moral assessment. More specifically, it clarifies at which stage the differences that we observe between groups in the OS are introduced.

We base our extension on the work of Rawls [22] who differentiates between *realized abilities* and *innate potential* (or, as Rawls writes, "native endowments"). Potential is innate to an individual and determined at birth.[1] This could, for example, be their innate intelligence or predisposition to develop those traits (e.g., extroversion) for being a good sales person.[2] The realized abilities represent how good they actually are at making sales at the time when the selectors are looking to hire. This may be influenced by early socialization in the family, the type of school they went to, the university they attended, the opportunities they were given (internships etc.) and so on. The realized abilities is what we will keep referring to as the CS in our extension. We introduce a new space which represents the potential: The PS.

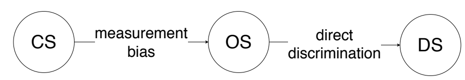
Figure 1 shows the *spaces* and *biases* that we differentiate in our model. The spaces can be understood as different stages: We start with our innate potential, represented by the PS, at birth. Shaped by our life experiences, we realize our abilities to potentially different degrees, which is captured in the CS. The realized abilities are then measured in the OS. The OS is used as the basis of the predictions in the DS.[3]

The introduction of the PS gives us the ability to differentiate between two types of "we're all equal", which are conflated in Friedler et al.'s description of WAE. We will define them as distinct worldviews. Note that besides assuming one of these worldviews, it is also possible to hold both views at the same time, or neither of them as they are not opposing.

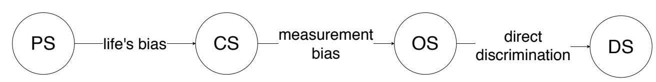
- "We're All Equal in the PS" (WAEPS), meaning all groups have the same innate potential. This means that lacking Life's Bias (LB) (which we shall define shortly) and MB all groups would be equal.
- "We're All Equal in the CS" (WAECS), meaning all groups are currently actually the same (even though it may look differently when taking measurements). This is the literal interpretation of "we're all equal" in which all groups are seen as already being equal today.

Note that the second type is the same as Yeom and Tschantz's definition of WAE. Formally, Friedler et al. also define WAE as equality in the CS, but leave the option that it may also be interpreted as equality in the PS.

The distinction between PS and CS allows us to define another type of bias. As already stated, Friedler et al. refer to the introduction of group differences from the CS to the OS as "structural bias" while we refer to it as MB since it is introduced through the act of measuring and is dependent on, e.g., availability of information or variable selection. As seen in Figure 1a, they also term the bias introducing group differences from the OS to the DS: direct discrimination. We introduce a third bias, which is the bias from the PS to the CS. Inequalities, such as differences in the qualities of schools and universities, the income and connections of their parents etc., can set individuals with the same potential far apart in terms of their realized abilities. We will refer to these inequalities as Life's Bias (LB). We remain neutral, at this stage, on whether LB is the same as injustice. We notice, in passing, that if injustice exists, it may affect LB. For example, if people routinely act based on gender stereotypes, men and women with the same potential may end up expressing different realized abilities to a different degree. Furthermore, if acting based on gender stereotypes is morally wrong (as it seems plausible), LB will be unjust. In cases like this, we shall refer to LB as *unjust* LB for precision's sake.



(a) Friedler et al.



(b) Our extension

Figure 1: Relationship between the spaces and biases.

3.3 Rules Distilled From the Framework

We will now again consider the framework proposed by Friedler et al. and its interpretation by Yeom and Tschantz to distill rules about when and why to choose independence as the fairness measure. The extension of Friedler et al.'s framework presented in the previous section will be used in order to clarify the position of the two papers.

We find that the papers consider the usage of independence from two perspective: One perspective specifies the assumptions justifying the enforcement of independence (`IF [assumption], THEN independence should be used`) and the other one describes which assumptions are implied when enforcing independence is argued for (`IF independence should be used, THEN [assumption]`). We will discuss both perspectives for both papers and derive propositions summarizing them as rules.

3.3.1 Friedler et al.'s rules. Let us first consider Friedler et al.'s framework. In order to understand their reasons for recommending independence, we have to introduce two key terms that appear in their proposal: Fairness and non-discrimination. We refer to [9] for a mathematically precise definition, but the idea can be expressed as follows:

- *Fairness*: Individuals who are close in the CS are (fairly) close in the DS.[4]
- *Non-discrimination*: The difference between groups is not (notably) increased from the CS to the DS.

[1] "At birth" here should be considered as "at conception" since there is already social influence at the fetal stage [12].

[2] "[N]ative endowments of various kinds (say, native intelligence or natural ability) are not fixed natural assets with a constant capacity. They are merely potential and cannot come to fruition apart form social conditions [...]. Educated and trained abilities are always a selection [...] from a wide range of possibilities that might have been fulfilled. Among what affects their realisation are social attitudes of encouragement and support, and institutions [...], opportunities and social position, and the influence of good and ill fortune." [22, pp. 56-57]

[3] Note that this resembles the proposition by Mitchell et al. [17] to differentiate between the "world as it should and could be", the "world as it is" and the "world according to data" [17, p. 4]. The "world as it is" is equivalent to our CS and the "world according to data" corresponds to the OS. It remains, however, unclear what the "world as it should and could be" really means.

[4] This corresponds to Dwork et al.'s *fairness constraint*, which requires that "similar individuals are treated similarly" [7].

In order to avoid confusion with the colloquial way of using these terms, we will avoid using these terms in any other way than defined by Friedler et al.. If we do use them in the colloquial or philosophical sense, we will make this evident from here on.

IF [assumption], THEN independence should be used. The first perspective asks the question what condition has to be met for suggesting the enforcement of independence. Friedler et al. state that "under a structural bias worldview, only group fairness mechanisms achieve non-discrimination (and individual fairness mechanisms are discriminatory)" [9, p. 12]. Note that "group fairness mechanisms" here refers to algorithms fulfilling independence (and not any group fairness metric) and that "structural bias" is what we refer to as MB. From this we can follow the rule that `IF there is MB`[5]`, THEN independence should be used` because we otherwise produce discrimination (i.e., the opposite of non-discrimination as described above).

Similarly, non-discrimination is also achieved through independence when Friedler et al.'s WAE is assumed: "when the we're all equal axiom is assumed group fairness mechanisms [i.e., independence] can be shown to guarantee non-discrimination" [9, p. 11]. As discussed above, Friedler et al.'s definition of WAE is equality in the CS, i.e., WAECS. The rule that follows is similar to the one above, namely `IF WAECS`[6]`, THEN independence should be used` because we might otherwise produce discrimination.

This leads to the following first rule for Friedler et al.:

Proposition 3.1. `IF there is MB OR`[7] `WAECS, THEN independence should be used.`

IF independence should be used, THEN [assumption]. Friedler et al. claim that "under a WYSIWYG worldview [i.e., no MB], [...] group fairness mechanisms [i.e., independence] are unfair" [9, p. 12]. This can be translated as `IF WYSIWYG, THEN NOT independence should be used` because we otherwise create unfairness. Since `WYSIWYG` is equivalent to `NOT there is MB`, the rule can be restated as `IF NOT there is MB, THEN NOT independence should be used`. From that we know that independence should only be used if there is MB. After all, if there is no MB, then independence should not be used. From this we can follow the other side of the rule: `IF independence should be used, THEN there is MB` because we otherwise create unfairness.

As with the first rule, there is another way of viewing the second rule. Friedler et al. write that when "the goal is to bring this difference [between groups in the DS] close to zero, the assumption is that groups should, as a whole, receive similar outcomes. This reflects an underlying assumption of the we're all equal axiom so that similar group outcomes will be non-discriminatory" [9, p. 14] Using independence thus reflects the assumption that WAE holds, so `IF independence should be used, THEN WAE`.

We follow the second part of the rule as follows:

Proposition 3.2. `IF independence should be used, THEN there is MB OR WAECS.`

[5]What we mean here is "we assume that there is MB". For brevity, however, we will simply write "there is MB" from now on.
[6]Again, we will write "WAECS" to say "we assume that WAECS".
[7]When using `OR`, we are referring to the logical operator $\lor$, which means that the statement is true if either one or both operands are true.

In this case, the mapping from the CS to the DS is again non-discriminatory. If WAE does not hold, the mapping would again be unfair.

Figure 2 shows these two rules and the implications of not following them.

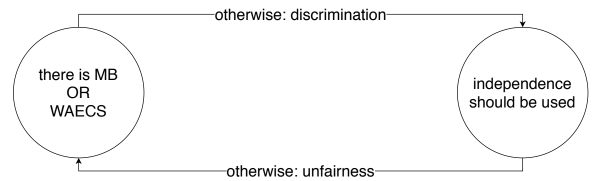


Figure 2: Friedler et al.'s rules for choosing independence.

3.3.2 *Yeom and Tschantz's rules.* Let us now consider the rules proposed by Yeom and Tschantz.

IF [assumption], THEN independence should be used. Yeom and Tschantz claim that "if the WAE worldview holds, we have $Y' \perp Z$ [i.e., independence between the CS and the groups], so every optimal model satisfies $\hat{Y} \perp Z$ [i.e., independence between the DS and the groups]. This implies demographic parity [i.e., independence]" [28, p. 5]. From this we follow the rule `IF WAECS, THEN independence should be used` because the model will otherwise not be optimal, meaning there will be a loss in utility.

Proposition 3.3. `IF WAECS, THEN independence should be used.`

IF independence should be used, THEN [assumption]. They further argue that "the We're All Equal (WAE) worldview motivates the demographic parity test, and if the worldview does not hold, the demographic parity test tends to lower the utility of the model" [28, p. 5]. Translated to our extension of the model, the first part of the statement is equivalent to the already found rule `IF WAECS, THEN independence should be used`. The second part of the rule can be translated as `IF NOT WAE, THEN NOT independence should be used` because we will otherwise lower the utility. The utility loss argument is quite intuitive: If there are inequalities in the realized abilities, i.e., in the qualifications of each group, then insisting on achieving independence means that candidates with lower qualifications will be selected. Assuming utility is measured as the performance of the employees on the job, we assume that this will lower the utility.

The found rule is equivalent to the following proposition:

Proposition 3.4. `IF independence should be used, THEN WAECS.`

Figure 3 combines the rules for Yeom and Tschantz.

3.3.3 *Comparison of the rules.* The rules proposed by Friedler et al. and Yeom and Tschantz are similar in their structure. While Yeom and Tschantz argues for/against independence in practical terms (lower utility), the harms described by Friedler et al. are statistical properties of the relationship between the CS and the DS (fairness and non-discrimination). We can merge both sets of rules into one as follows:

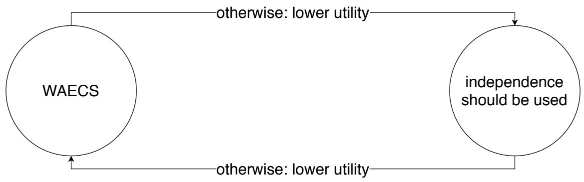


Figure 3: Yeom and Tschantz's rules for choosing independence.

PROPOSITION 3.5. `IFF (there is MB OR WAECS), THEN independence should be used`.

Note that `IFF x, THEN y` is equivalent to saying `IF x, THEN y AND IF y THEN x`. Proposition 3.5 is therefore simply a shorter version of combining Proposition 3.1 and Proposition 3.2, which were derived from Friedler et al.. Proposition 3.5 also contains Proposition 3.3 and Proposition 3.4 as it includes the WAECS assumption of Yeom and Tschantz. The only difference to the previously defined propositions is that we now not only argue for/against independence because of *either* the statistical properties, fairness and non-discrimination, *or* utility, but because of *both*. This can be seen when comparing Figure 4 with Figures 2 and 3.

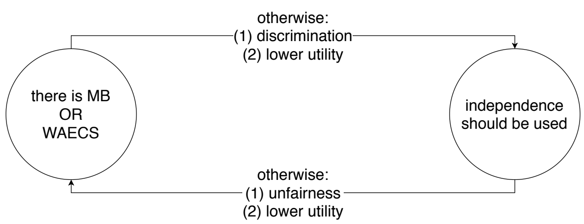


Figure 4: Merged rules for choosing independence.

Note though that both reasons – statistical properties and utility – are seemingly purely mathematical, not moral. However, Friedler et al. still use philosophical terms (fairness and non-discrimination) to reason for/against the enforcement of independence. Lipton and Steinhardt refers to the naming of technical concepts with terms that are colloquially used as *suggestive definitions*: These terms carry meaning in day-to-day life and therefore imply that our intuitive understanding of these terms in some way aligns with their technical definition [13]. This raises the question whether these terms aim at not only providing a purely statistical reasoning about independence, but also a moral one. As demonstrated above, if we interpret the papers literally, then they only provide mathematical justifications for the introduced propositions. However, since the DS includes predictions that are used to make potentially life-altering decisions, it is hard to see how the propositions could only be concerned with the mathematical validity of the predictions without any further moral considerations. In the next section, we will therefore consider the natural extension of the propositions: We will view them as providing moral reasons for/against choosing independence as the correct fairness metric. In doing so, we must consider all spaces introduced in Section 3.2. The moral rules that we introduce therefore include not only WAECS and MB, but also WAEPS and LB.

4 DEALING WITH LB

As we have shown, the literal interpretation of the WAE worldview states that all groups are actually the same in the CS, even though it may look differently when taking measurements. We should, however, consider what follows from assuming the WAE worldview at the PS level, while allowing for the existence of LB. Clearly, the result may be that we are no longer all equal in the CS since LB may affect members of different groups in different ways.

4.1 Two Ways of Dealing With LB in Distilled Rules

We will now discuss how LB may be included in the rules proposed by Friedler et al. and Yeom and Tschantz. The first way of dealing with LB in the rules distilled from the literature is to preserve the rules and simply apply them the same way when there is LB. This direction prioritizes Friedler et al. and Yeom and Tschantz's original reasoning about statistical properties and utility. These properties are what "fairness" (in its moral sense) ultimately reduces to when dealing with LB this way. The second way of including LB in the rules is to extend them and make the argument a moral one. For this, we extend the rules to the new type of bias, LB, by arguing by analogy: one could reason that LB – in terms of its moral features – is sufficiently similar to MB, so that similar rules follow when there is LB instead of MB or both. Indeed, the analogy between MB and LB will be stronger if we can identify a common moral principle that applies to both cases.

Let us now examine both ways for dealing with LB in turn.

4.1.1 The mathematical and utility-based extension. Proposition 3.5 states that `IFF (there is MB OR WAECS), THEN independence should be used`. As mentioned above, we are interested in the case where WAEPS and inequalities in the CS are caused by LB. This means that by hypothesis `NOT WAECS`. It follows from Proposition 3.5 that independence should not be used in this case. The reasoning is the same as given by Friedler et al. and Yeom and Tschantz: If WAECS is not assumed, enforcing independence violates Friedler et al.'s statistical fairness property (that individuals close in the CS space should be close in the DS) – independently of whether the PS is equal or unequal. Moreover, ignoring inequality in the CS space, again independently of whether the PS is equal or unequal, leads to a utility loss. This line of argument may, however, leave scholars that are interested in fairness (in its moral sense, not Friedler et al.'s statistical concept) dissatisfied.

4.1.2 The fairness-based extension. Let us now see why the mathematical and utility-based extension may be seen as insufficient and what the common moral principle between MB and LB is that allows us to include LB in Proposition 3.5. We will first formulate the argument in a way that should be easily accessible for non-philosophers. We then make the same argument in a philosophically more rigorous way. (Readers not familiar with philosophy can easily skip that section and still follow the argument.)

Accessible formulation. One may postulate the existence of two kinds of fairness scholars: those who are interested in fairness (in its moral sense) *only as long as* it aligns with utility, and those who are authentically interested in fairness (in its moral sense), i.e., interested in fairness *for its own sake*, also when it conflicts with utility. A scholar authentically interested in fairness may question the conclusion of the rules while acknowledging that from the point of view of utility independence should not be chosen. They may point out that, from the point of view of fairness, the situation in which MB causes inequalities in the OS that are not grounded in the CS (which justifies independence) is intuitively analogous to the situation in which LB causes inequalities in the CS that are not grounded in the PS. For clearly, they may argue, individuals who are equal in their PS do not deserve to be unequally ranked in the CS space any more than individuals who are equal in the CS space deserve to be unequally ranked in the OS.

Philosophically rigorous formulation. A more philosophical way of phrasing the previous point is as follows: Proposition 3.5 can be given a consequentialist and (prima facie) prudential justification, based on utility considerations; or it can be given a non-purely consequentialist and more distinctively fairness-based one,[8] based on considerations of *merit* or *responsibility*. The first argument considers the decision-maker as an agent who should make the prudentially best choice, i.e., maximize its utility. It therefore identifies the real-world circumstances in which using independence may align with utility maximization for the decision maker, and those in which it does not. (That means that if WAECS, but MB exists, letting MB be the ground of the decision does not promote utility [28].) This argument is *formally* prudential, that is, self-interested, because it asks what maximizes the utility of the decision-maker, not (or only derivatively) that of the people affected by its decisions.[9]

The more distinctive fairness-based justification for Proposition 3.5 is the moral intuition that it is *unjust* (and not merely, inefficient) for a decision about the individual to be taken when MB exists. There are at least two possible philosophical analyses of this fairness intuition, a merit-based one, and a responsibility-based one. Consider merit first: The conventional view of merit is that it is based on what people do [16, 19], for example, their contribution to society. Suppose that the actual contribution to society of two people, A and B, is equal, that is A and B have the same CS features. However, MB exists, so A is perceived to contribute more and, consequently, A receives a benefit that is denied to B. This is intuitively unfair, since A does not deserve a favorable treatment compared to B. Consider responsibility now: When MB exists, the individuals who are judged in a biased manner do not control the bias in the eyes/algorithm of the observer, and are not morally responsible for it. Intuitively, it is unfair that individuals are imposed costs due to factors for which they are not responsible. Summing up, the moral view behind proposition Proposition 3.5 is that if WAECS is assumed, then differences in the DS are unjust in the sense that they are undeserved (or on responsibility grounds).

Conclusion of the fairness-based extension. This suggests the following view as the natural extension of Proposition 3.5, i.e., the view that `IFF (there is MB OR WAECS), THEN independence should be used`:

Proposition 4.1. `IFF WAEPS AND (there is MB OR there is LB), THEN independence should be used`

4.2 Not All LB Should Be Corrected

Intuitively, Proposition 4.1 states that independence is called for not only to correct for MB, but also to correct for LB. Notice that Proposition 4.1 is now arguably too broad in the inequalities it promises to correct for. The fairness-based argument for removing the influence of MB was that it was neither morally neutral, nor merely inefficient, but actually *unjust* when it does not reflect merit or responsibility. Something similar, intuitively, must hold in this case. Namely, it is *unjust* LB that calls for some kind of correction or rectification.

4.2.1 Distinction between just and unjust LB. Note that it is a logical possibility that indeed all LB, *as such*, is unjust. If so, there is no distinction between *unjust* LB and LB simpliciter. This is the position that no one deserves the values in the CS which are influenced by any type of LB. It is, however, also possible to maintain a more nuanced view. It is easy to show this by considering theories of justice that political philosophers *actually* defend. Different substantive theories of justice provide different (and often irreconcilable) criteria for evaluating the justice of social structures. For example, institutional luck-egalitarianism maintains that all inequalities for which individuals are not responsible are unjust. Unjust inequalities are the ones which could have been prevented or redressed through suitable and feasible institutional arrangements. Rawls's theory of justice, on the other hand, maintains that inequalities reflecting people's unequal native endowments and motivations are just – provided that (1) they are not influenced by the social class of birth and (2) emerge through institutions arranged in a way that delivers the greatest expectations of social primary goods to society's least advantaged members.

These theories (and many others) disagree when arguing about the justice of institutions. The luck-egalitarian one, for example, commits the government to do everything it can to level the playing field among individuals born with unequal natural endowments, as these inequalities are undeserved. Rawls's view, however, approves of such inequalities if they boost productivity in a society where people at the bottom of the social pyramid are the ones to benefit the most from such productivity gains. Yet these views also converge on many real world cases: for example, current US society is arguably very unjust according to both views.

Arguably, from the perspective of both luck egalitarianism and Rawls's theory of justice, many inequalities in the CS, which are produced by LB, actually reflect unjust social structures. They are instances of *unjust* LB. Notice that neither theory implies that all LB is unjust LB. Let us consider the luck-egalitarian view that only factors for which one can be held morally responsible justify inequality. It may still be objected that the development of innate potential into realized abilities is not determined entirely by external circumstances that are matters of brute (good or bad) luck. Apart from the influence of good and ill fortune, our realized abilities reflect our personal history, i.e., the choices we make, every day. If human agency is not an illusion, we are (partially, at least) responsible

[8] Assuming that fairness is not reducible to utility maximization, as all non-utilitarians believe.

[9] The designation of "prudential" is, however, only *prima facie* since, in many cases, the self-interest of the decision-maker is supposed to serve the well-being of society as a whole, not the individual well-being of the decision maker. For example, a parole officer is supposed to make decisions promoting the interest of society to prevent crime while allowing the gradual inclusion of the offender who has almost served his sentence in society, not to make the decision that serves his own personal career goal.

for at least some of our choices. Therefore, at least some LB is not morally problematic. It is therefore unclear why one should treat inequalities arising in the CS as a result of such LB on a par with inequalities arising in the OS due to MB.

4.2.2 Examples of just LB. In the discussion above, we are saying that not all LB is necessarily unjust because some LB might just be caused by personal choices. However, one may question whether it is possible for personal choices to cause unequal outcomes, not only individually, but also at the group level, even if there were no inequalities in the PS. For example, two people with the same potential, with institutions that only let inequalities reflecting their individual choices exist, could still end up with different realized abilities. Yet, on a group level, it seems improbable that one group justly has a statistical prevalence of individuals making one kind of choices, for which they can be held responsible, and another group justly has a statistical prevalence of individuals making a different kind of choice. If that different prevalence exists and we assume WAEPS, certainly there must be something causing the group inequality for which individuals cannot and shall not be held responsible. Therefore, at least for a luck-egalitarian view of injustice, the emerging LB is also necessarily unjust. We actually believe that this argument, while prima facie plausible, is not entirely correct, as it fails to account for some special counterexamples.[10]

Beside these not-so-intuitive examples, we should also consider moral views that differ from luck-egalitarianism. For example, one may hold that the influence of parents who read bedtime stories to their children and in this way cause unequal IQ leading to unequal outcome is never unjust [14], a view that contradicts luck-egalitarianism [25]. If that view is correct, inequalities that have been created by reading bedtimes to children are not unjust even if it so happens that, for historical and cultural reasons, reading bedtime stories to children is more habitual in certain cultures than others.

When LB exists, but is not unjust, the moral reason for enforcing independence no longer holds. In other words, the decision whether independence should be used depends not only on facts but also on values when LB exists. According to luck egalitarianism, independence is not required, if the following two requirements are fulfilled: (1) unequal decisions (i.e., inequality in the DS) emerge purely as a result of unbiased observations (i.e., WYSIWYG) of the features that we *want* to base the decision on (i.e., CS) and (2) these features are unequally distributed (in spite of equality of potential in the PS) simply as a result of choices for which individuals can be considered *fully responsible* for. According to other theories of justice, independence is not required, if the features in question emerge as a result of processes such as reading bedtime stories to one's children.

[10]Suppose that the oppressed group in society finds a new leader who starts a revolution and achieves equality for all. The members of the former oppressed group develop special skills deriving from their fight against the oppressor. Suppose that these skills are observed and appropriately valued by employers in the post-revolutionary world. If we now assume that the revolution is just and makes up for the injustice of the oppression, then the differences in the OS are cause by the LB, but this LB is seen as just (because revolutionaries are rightly held responsible for their actions as revolutionaries). In this case, the WYSIWYG worldview seems to be appropriate even if independence relative to the group of origin is violated because individuals who have freed themselves from the chains of oppression deserve the advantageous abilities that such struggle has brought to them.

4.3 Extended Rules: Final Formulation

As we have shown, the theory of justice one adheres to determines one's judgment as to whether LB provides a reason to enforce independence, or not. In conclusion: the most charitable interpretation of the extension of Proposition 3.5 to include LB is not Proposition 4.1, above, but rather, the following more nuanced view.

Proposition 4.2. `IF WAEPS AND (there is MB OR there is unjust LB), THEN independence should be used`

In what follows we will refer to the underlying assumptions as the "We're All Equal But There Is Injustice" (WAEBI) worldview.

Definition 4.1 (WAEBI worldview)**.** The WAEBI worldview subsumes the following assumptions:

(1) WAEPS and
(2) there is unjust LB.[11]

Let us now come to a synthesis by combining the argument that `IF WAEBI, THEN independence should be used` with the previous arguments associating independence with the WAE view. As already discussed, the WAE axiom fails to distinguish between WAECS and WAEPS. Equivalently, the MB worldview fails to distinguish between MB and LB. We have argued that if we consider LB, we should consider a moralized version of the relation between LB and independence (i.e., not all forms of LB require to be corrected by enforcing independence). The overall view is Proposition 4.2. Let us focus on one direction of Proposition 4.2: `IF (there is MB OR there is unjust LB), THEN independence should be used`.

Taking this into consideration, we will consider the view Proposition 4.2 in what follows as a natural extension of Proposition 3.5. However, we want to focus our attention on the implications of Proposition 4.2 on the scenario in which MB does not exist in order to simplify our discussion somewhat. Proposition 4.3 follows from Proposition 4.2 if it is assumed that there is no MB.

Proposition 4.3. `IFF WAEBI, THEN independence should be used`

5 TWO COUNTEREXAMPLES AGAINST EXTENDED RULES

From this point of the paper on, we focus on unjust LB, which calls for an engagement with moral/political philosophy more than MB does. Arguably, this is because:

(1) The question how social institutions affect LB is recognized by some of the most influential scholars of political philosophy as the subject matter of "A Theory of Justice" [22]. John Rawls, in particular, identifies social justice – a subject matter of political philosophy – as the question pertaining to the justice of the basic structure of society. Rawls's definition of the basic structure picks out those institutions that are apt to cause LB as we define it [22]. So LB is definitely a topic about which political philosophy has something to say.
(2) WAEBI is a value-laden condition, as already in the definition it mentions the moral concept of injustice. By contrast, MB

[11]Note that this second assumption logically leads to the assumption of inequalities in the CS, which are unjust.

only involves difference between what actually exists (in the CS) and what is observed (in the OS), which can also be described in a morally neutral language.

As stated before, we focus on proposition Proposition 4.3, which simplifies the discussion of Proposition 4.2. Proposition 4.3 is the claim that `IFF WAEBI, THEN independence should be used`.

Clearly, when there is no MB, Proposition 4.3 is true if and only if Proposition 4.2 is true. Thus, we can simplify the analysis of Proposition 4.2 somewhat by focusing on Proposition 4.3, assuming that no MB exists. We shall proceed in a logical fashion, by investigating the two parts of the biconditional in turn:

(1) `IF WAEBI, THEN independence should be used` and
(2) `IF independence should be used, THEN WAEBI`

Notice that rule 2 is informative, even though we already know that it is incorrect because independence should also be used when MB exists, even if WAEBI is not assumed. The reason why it is still informative is that we will show that the claim is incorrect – independently of MB. We argue against both rule 1 and 2 by counterexample. The counterexample to rule 2 shows not only that Proposition 4.3 is incorrect, but also that Proposition 4.2 is since the counterexamples do not involve MB. Thus, we refer to Proposition 4.3, so that we can bracket the issue of MB and avoid distractions.

5.1 Counterexample Against Rule 1: IF independence should be used, THEN WAEBI

It is now time to offer a convincing counterexample to the view that (absent MB) independence is required only if WAEBI.

For this, recall the definition of WAEBI, Definition 4.1, which states that this worldview assumes WAEPS and unjust LB. The counterexample is a case in which the WAEBI assumptions are all satisfied, yet independence is not required. One can build a counterexample as follows:

(1) First, let us suppose that there is a specific severe congenital disorder that is very painful and drastically reduces the individual's life expectancy. We will refer to this specific severe congenital disorder as *SCD*. Let us further assume that (probably contrary to fact) all individuals are generally equally at risk of being born with *SCD*. WAEPS is therefore satisfied.
(2) Second, let us suppose that – while the risk for being born with *SCD* is generally the same for all individuals – this risk is notably increased when the mother breathes in dangerous pollutants during pregnancy. Assume now that mothers in one group, e.g. the *green* group, are more likely to live in neighborhoods close to chemical factories that emit dangerous pollutants. Individuals in the green group are thus more likely to be exposed to the risk of developing *SCD*. We shall suppose that this unequal exposure is produced by huge and uncontroversial injustices in society. *Green* mothers might, for example, be more likely to live in poverty because of direct discrimination against them, which makes their opportunities for all sorts of job worse than those of the *orange* group. For this reason, they cannot afford moving and have to live in poor neighborhoods plagued by dangerous pollutants. (This case plausibly counts as injustice even according to more moderate forms of egalitarianism than luck egalitarianism.) Hence, there is unjust LB.
(3) Third, as a result of 2, members of the *green* group are more likely to suffer from *SCD* than members of the *orange* group. Thus, there is an unjust inequality in the CS. (We shall see next why *SCD* is the CS.) We shall suppose that whether a patient suffers from *SCD* is a clear cut, binary condition, i.e., either someone does, or does not. There are no intermediate stages.

Suppose that a very expensive therapy is developed, which cures people with *SCD* but causes recurrent migraine (with moderate frequency, let us say, one per month). As *SCD* has bad consequences for the individual (pain, drastically shortened life), we shall assume that the benefits of the therapy outweigh its costs. Suppose now that the therapy only works if it affects fetal development. Thus, in order to avoid the disease for the future individual, it is the mother that has to be treated before the illness is fully manifested in the child.

Suppose that machine learning specialists develop a perfect accuracy predictor to determine, based on a non-invasive clinical examination, whether the fetus will be ill. (This may be impossible in practice. However, in a philosophical argument, we can test the theory also with hypothetical examples. The challenge is to explain what could be morally wrong with the independence-fulfilling predictor. Notice also that in the clinical setting one can already make high accuracy predictions. With close to perfect accuracy, people often act and reason as if the accuracy was perfect.) Since the predictor is perfectly accurate, it will predict *SCD* at a higher rate for the green than for the orange patients. As a result, green patients will receive the therapy more often than orange patients do, which violates independence.

We will now argue that this perfect accuracy predictor is perfectly just. The argument we present is very robust because it is coherent with ethical views that sometimes pull in different directions and, intuitively, it is difficult to make the case that the argument is wrong. Indeed, it should be so obvious that the predictor is fair, that it would be counted against any view entailing the opposite for this case, that it cannot align with this result. The decision of the perfect predictor is perfectly fair because no individual has a claim against the distribution based on it. By "no individual has a claim", we do not mean that some individual may have a *prima facie* claim that a different decision should be taken, which is then overridden by the claims of others. We also do not mean that some individual has a claim that holds *prima facie*, but that is defeated by some substantive view of justice, which the individual, if reasonable or endowed with moral sensibility, should respect (even if it is not in the individual's own interest to respect it). What we actually mean by "no individual has a claim" is the much more radical claim that the individual has no claim against the perfectly accurate distribution (in this case), not even a *pro tanto* or a *prima facie* claim.

To see why no individual has a claim against the perfect accuracy distribution, consider that no individual, faced with the decision by a perfect predictor, can point to an alternative distribution that they have any reason to prefer. This clearly is the case in the example.[12] For, first, each individual person who will develop *SCD* is better off with a decision based on the correct prediction because the individual is certain to receive the cure, which is the preferable outcome despite the side effects. Second, every individual who will not develop *SCD* is better off without the therapy because the individual is certain to not need the cure. Not receiving the therapy is thus the preferable outcome as it avoids the side effects. As a consequence, no one has a claim to a different decision. Moreover, any departure from the perfect accuracy predictor makes someone worse off and no one better off. When the features in the CS are not equally distributed between the two groups (i.e., *green* and *orange*, in this case), enforcing independence sacrifices some accuracy. Suppose that this sacrifice amounts to a single wrong diagnosis. That means: either someone who will actually develop *SCD* will not receive the cure, or someone who will not develop *SCD* will receive the cure. Either way, the choice to enforce independence will cause harm to at least one individual, which gives that individual a claim *against* independence. It seems that this is one rare case in which one view of what is fair is truly robust because, besides maximizing utility, no individual has a claim against the perfect accuracy predictor, even if it violates independence. Furthermore, if independence is enforced in this case even though it causes inaccuracy, there will be at least one individual who has a moral claim against independence being enforced. This claim entails that enforcing independence is morally wrong because it is not defeated by any claim *in favor of* enforcing independence. The question of comparing the relative urgency or strength of moral claims does not even arise.

Our argument here is not merely that independence in this case involves a loss of accuracy (and thus utility) and that, simply for that reason, is the morally wrong choice in this case. While it is correct that there is a conflict between independence and accuracy in this case, our argument is much stronger than the usual utilitarian argument. The usual utilitarian argument points out that enforcing independence causes a loss of *aggregate* utility [6]. This argument also focuses on utility, but it considers it from the perspective of each and every individual involved in the decision. A utilitarian argument would object that enforcing independence causes a utility loss in the aggregate and that for that reason it should not be done. However, such an argument also requires that, in order to reach the utilitarian maximum, some people are made worse off for the benefit of other people. The utilitarian view is that this is always morally right when the sum of utility is maximized. Many people find this view objectionable (e.g., [22]). The objection against the utilitarian is that it does not respect the *separateness of persons* [22]. Our argument against independence does not imply the utilitarian conclusion, so it is is not vulnerable to that objection.[13]

Summing up, it is not true that `IF WAEBI, THEN independence should be used`. In this case, WAEBI is clearly satisfied (by hypothesis), and yet independence should not be used.

[12]It may be objected that this is a very peculiar example, and that not all cases involving perfect accuracy predictors are relevantly similar. That is probably correct. However, one case is all it takes to generate a counterexample that falsifies a general claim about when independence should be used.

[13]Our argument is a contractualist one, not a utilitarian one [23]. Our thesis that the perfect accuracy predictor is just is so robust because it is independently supported by contractualism and utilitarianism.

5.2 Counterexample Against Rule 2: IF independence should be used, THEN WAEBI

Now let us turn to the other direction of the biconditional, which is the idea that `IF independence should be used, THEN WAEBI` is assumed. A counterexample to this would be a case in which, independence seems intuitively fair or called for, yet WAEBI conditions are not satisfied. Unfortunately, this example is not as robust as the first one is. The example itself is inspired by a fairness theory for machine learning inspired by economic and political theories of equality of opportunity, which provides indications for when independence should be used [10]. We do not rely on this theory, as we find that a strong case can be made for the conclusion on intuitive grounds.

We consider the design of an algorithmic decision system deployed after natural disasters. This decision system is tasked with determining where drones should be sent in order to attempt to rescue civilians from drowning after their houses and streets have been flooded. Data scientists train a machine learning model to decide where to send the drones in such cases. The initial goal is to simply maximize the number of lives saved.

Let assume that there is a flooding which affects a city with its surrounding suburbs. While the city is densely populated, the suburbs are not. We can split the population into two demographic groups: the *green* and the *orange* group. It turns out that the *orange* group tends to live in the city and the *green* group tends to live in the suburbs. Because of the difference in population density between the city and the suburbs, a drone that is sent to the city has a much higher base rate probability of resulting in a successful rescue. Hence, the utility-maximizing model is more likely to send drones to the densely populated city than to the suburbs – it diverts resources to the suburbs only when a large proportion in the cities has been saved. As a consequence, the probability to be saved is much higher (say, ten times higher) if you are *orange*. This means that members of the *green* population are very unlikely to be rescued.[14] We maintain that in this case independence is morally required. The reasoning is the following: Every individual equally needs to be saved, independently of where they live, and no one should be held morally responsible for failing to live in a relatively densely populated area, for matters of life and death. Thus, in a sense, everyone equally deserves to be saved. If everyone equally deserves to be rescued, everyone should have the same prospects of being rescued, independently of their sex, race, or any other sensitive attribute. If so, any inequality in the probability of rescue associated with membership to a group is morally problematic, for it cannot be justified based on merit, or need, or responsibility.

It may be objected that there is a clear moral reason to prioritize saving urban individuals, namely that this will maximize the total number of lives rescued (and we ought to maximize this value). However, notice again, that this is a utilitarian, maximizing argument. Most moral problems of fairness in machine learning, or at least most *morally deep* problems, emerge because there is a conflict

[14]Clearly we assume here that data scientists cannot reach, or even approximate, a perfect accuracy predictor. This implies that when the algorithm predicts that a person will be saved by sending a helicopter to coordinates X, Y, Z, it is not always the case that someone will get rescued, particularly in the suburbs.

between maximizing utility and fairness (in its moral sense) defined in a way that is independent from it. Hence, in a sense, the fact that the fairness intuition conflicts with a utilitarian assessment of what should be done is to be expected for an authentic (non utilitarian) moral intuition for fairness. Arguably, the best way to take the utilitarian intuition – that there is a (moral) reason to send drones predominantly into the densely populated areas – into account is by viewing it as a consideration of efficiency that an ethically sound procedure should balance with considerations of distributive justice. That is, the *all-things-considered* morally desirable algorithm will neither be one that maximizes utility, nor one that achieves independence fully, but rather something in between, that will compromise utility, to some extent, but also achieve a more balanced rescue of the two populations. We then conclude that this is a case in which independence should be used due to a concern with fairness (in its moral sense). (Even if, let us grant the objection, not fully achieved as fairness needs to be balanced with efficiency.)

Having argued that independence should be used, this amounts to the counterexample we are looking for if we show that the moral case for independence is independent of the WAEBI conditions. For building the stronger possible case, we shall suppose that *every single one* of the conditions that jointly define the WAEBI worldview is false.

First, we do not assume that WAEPS, that is, people are born with a disposition to live in cities or suburbs. For example, some people live in the city just because they are born there, even if it is not true of everyone.

Second, it is not the case that LB exists, or at least, the plausibility of the conclusion about fairness does not depend on the existence of LB. We may consider, for the sake of the argument, a society in which people are not pressured to live in cities. The case for rescuing the people in the suburbs is as strong in a society in which people are not pressured to live in cities, as it is in one in which they are pressured to do so (among other things, by the perception that their lives have less value in case of rescue, if they remain in less densely populated areas). So the conclusion does not depend on the existence of LB.

Third, we may as well suppose that there is LB, but the LB is not *unjust*. For example, people end up living in suburbs and cities (and different groups, e.g., *green* and *orange*, have different propensities to do so), but this is not in itself unjust or the result of injustice in society. Schelling's model of segregation shows that a mild preference for living among members of the same group will over time lead to segregation [24]. For this example, we assume that the *green* and *orange* population have a slight preference for members of their own group and that this preference is innate and not caused by injustices. Over time the two groups have segregated to some extent, so that the majority of the people living in the city happens to be *orange* and the majority of the people living in the suburbs is *green*. Segregation observed in the real world is probably at least partially caused by this in-group preference. However, we also note that there are of course unjust reasons for segregation. We in particular point out the US context where whites are more likely to inhabit less densely populated areas than blacks. The current levels of segregation in housing may be seen as the result of injustices such as historically forced segregation [21], exclusionary zoning [26], white flight [8], ongoing discrimination and racial steering [5] as well as seemingly race-neutral laws that encourage segregation [2].

In conclusion, we have identified a case in which independence should (plausibly) be used. And yet, in this case, the conditions realizing the WAEBI worldview are not satisfied. This counts as a counterexample to the claim that `IF independence should be used, THEN WAEBI`, and concludes our rebuttal of the biconditional claim Proposition 4.3. Since neither example depends on the existence of MB, the arguments also disprove Proposition 4.2.

6 CONCLUSION

In this paper, we have analyzed one argument that can be given in support of enforcing independence in a machine learning model, found in the recent machine learning literature. This argument claims that one shall enforce independence (i.e., use it as a fairness constraint of the model) if (and only if) "We're All Equal" (WAE) or there is Measurement Bias (MB). We have introduced the concept of Life's Bias (LB) as a type of bias that can be distinguished from the MB proposed by Friedler et al. (They call this type of bias "structural bias".) This shows that the WAE view as stated in the literature is incomplete.

We have identified two possible extensions of the argument presented in the literature, which are relevant when inequalities are generated by LB. We argue that the most (morally) plausible extension of the view discussed in the WAE literature is the view that one should enforce independence if (and only if) "We're All Equal But There Is Injustice" (WAEBI) or if there is MB. In other words, the justification of independence depends on the truth of the WAEBI worldview, which assumes that socio-demographic groups have similar innate potential at birth, but unjust LB leads to differences in their realized abilities.

Unfortunately, we found two powerful counterexamples to this ideally simple view: one clearly showing that injustice does not always morally require enforcing independence, the other making it plausible that (even in the absence of MB) injustice is not morally required for the use of independence to be morally justified.

The conclusion is that the relation between independence and the WAEBI view is not simple at all. The relatively simple and morally plausible proposition linking WAEBI and independence we analyze here is not universally true.

This is an important negative finding for the ethics of choosing fairness definitions for machine learning. Future research should either find an alternative framework to justify using independence as a fairness constraint in machine learning, or qualify the WAEBI view (in a principled way) to immunize it from such counterexamples. Both avenues seem very challenging.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2016).

[2] Deborah N Archer. 2019. The Housing Segregation: The Jim Crow Effects of Crime-Free Housing Ordinances. *Mich. L. Rev.* 118 (2019), 173.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2020. Fairness and Machine Learning.

[4] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[5] Camille Zubrinsky Charles. 2003. The dynamics of racial residential segregation. *Annual review of sociology* 29, 1 (2003), 167–207.
[6] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.
[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[8] William H Frey. 1979. Central city white flight: Racial and nonracial causes. *American Sociological Review* (1979), 425–448.
[9] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).
[10] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. 2019. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 181–190.
[11] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[12] Eszter Kollar and Michele Loi. 2015. Prenatal equality of opportunity. *Journal of Applied Philosophy* 32, 1 (2015), 35–49.
[13] Zachary C Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341* (2018).
[14] Andrew Mason. 2006. *Levelling the playing field: The idea of equal opportunity and its place in egalitarian thought*. Oxford University Press.
[15] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
[16] David Miller. 1979. *Social justice*. OUP Oxford.
[17] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).
[18] Thomas Mulligan. 2017. Uncertainty in hiring does not justify affirmative action. *Philosophia* 45, 3 (2017), 1299–1311.
[19] Serena Olsaretti. 2006. Justice, luck, and desert. *The Oxford handbook of political theory* (2006), 436–449.
[20] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.
[21] Garrett Power. 1983. Apartheid Baltimore style: The residential segregation ordinances of 1910-1913. *Md. L. Rev.* 42 (1983), 289.
[22] John Rawls. 2001. *Justice as fairness: A restatement*. Harvard University Press.
[23] Thomas Scanlon. 2000. *What we owe to each other*. Belknap Press.
[24] Thomas C Schelling. 1971. Dynamic models of segregation. *Journal of mathematical sociology* 1, 2 (1971), 143–186.
[25] Shlomi Segall. 2011. If you're a luck egalitarian, how come you read bedtime stories to your children? *Critical Review of International Social and Political Philosophy* 14, 1 (2011), 23–40.
[26] Christopher Silver. 1991. The racial origins of zoning: Southern cities from 1910–40. *Planning Perspective* 6, 2 (1991), 189–205.
[27] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[28] Samuel Yeom and Michael Carl Tschantz. 2018. Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. *arXiv preprint arXiv:1808.08619* (2018).