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 proposed, 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. high school grades and standardized test scores are predictive of performance at university) between socio-demographic groups are caused by unjust social disparities or 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.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → Machine learning; • **Social and professional topics** → Socio-technical systems.

KEYWORDS

fairness, independence, statistical parity, distributive justice, bias

ACM Reference Format:
Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the Moral Justification of Statistical Parity. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442188.3445936

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445936

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 straightforward as computing statistical metrics. It might require discussions with the stakeholders of the application and finding a compromise – after all, there are very few cases where everyone can agree on the correct choice of fairness metrics, in particular because it has been shown that some of them are conflicting [4, 9, 20]. It is thus important to not only provide mathematical definitions of fairness metrics, but to provide some guidance on 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 from the birth of an individual to the point where their abilities are measured.

2 WHAT IS INDEPENDENCE?

The arguably most commonly used category of fairness metrics is referred to as *group fairness* and focuses on the question whether socio-demographic groups are treated similarly or receive similar outcomes [12, 24]. 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* [4].

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 $P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A \neq a)$ [38]. 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: $R \perp A$ [4, 26]. This general proposition is referred to as the fairness criterion *independence* [4]. Independence can be found implemented in practice, in particular in the HR domain[1], and has been particularly influential in the early stages of the algorithmic fairness literature (see, e.g., [7, 29]). Together with independence, [4] lists *separation* and *sufficiency* as the three fairness criteria that most fairness metrics that have been proposed in the literature are closely related to. Separation and sufficiency gained more attention in recent years through the debate sparked by Angwin et al.'s investigative article *Machine Bias* [2]. 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 [4, 9, 20]. 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 enforcing independence.

Unsurprisingly, machine learning models trained to optimize, for example, accuracy, rarely coincidentally fulfill independence. However, we can enforce independence through various strategies (see, e.g., [7, 19]). 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 utilitarian moral goals such as maximizing 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. [16] propose a framework which enables its users to clarify the *worldview* assumed in the context of their application. This section will discuss Friedler et al.'s paper and distill the implicit rules for when to apply independence.[2] 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. This is reflected in the main result of [16], which is the distinction between the following three spaces (see Figure 1a):

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

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 [16]: hiring. In this case, the task is to predict employee productivity and take a hiring decision based on this prediction. We borrow the language for this example from [27]. A company picks whom to hire from the pool of *candidates*, i.e., the people who apply for the job. In order to make this decision, the company tries to predict who will perform best when hired. Each candidate brings certain *qualifications*, based on which the company wishes to make its 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, the company only has 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 company only has access to the qualifications through the proxies, it has to base its hiring decisions on the proxies. In this example, the qualifications are equivalent to the unobservable CS while the proxies represent 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. This implies that observed differences between sociodemographic groups correspond to actual differences. In our example, that would mean that the usage of CVs, interviews

[1] The popularity of independence in HR is attributable to the notion of *disparate impact* found in the Uniform Guidelines on Employment Selection Procedures introduced by the Equal Employment Opportunity Commission (EEOC) in 1978 [14]. Disparate impact means that one group is disproportionally affected by e.g. a hiring policy. When bringing forth a disparate impact claim, the "4/5ths" rule works as a rule of thumb: It compares the share of hired people from each group. If one group's hiring rate is less than 4/5ths of the hiring rate of the other group, then HR might be liable for disparate impact discrimination [14]. This 4/5ths rule is essentially an expression of statistical parity that allows for some leeway: Instead of enforcing the perfect equalization of hiring rates, it allows for some difference in hiring rates. Even though this rule is just a rule of thumb, Raghavan et al. [31] recently showed that vendors of algorithmic hiring tools treat it as a hard constraint. The web demo of AIF360 [5], a tool built for the bias auditing of predictive models, refers to statistical parity as being fulfilled if the 4/5ths rule is fulfilled. We thus see that through the 4/5ths rule, the notion of independence is highly relevant in practice.

[2] We discuss this particular paper as it uncovers the hidden assumptions that seem to be held when algorithmic fairness scholars advocate for independence [1]. The paper is well-known in the field and has influenced both theoretical and practical work. On the practical side, the two opposing worldviews have been described in AIF360's web demo to guide practitioners in their choice of fairness criterion [5]. Theoretical work has built on Friedler et al.'s framework to, e.g., provide mathematical ways to interpolate the proposed opposing worldviews [17, 40]. Mitchell et al. [26] cite Friedler et al.'s reasoning when explaining when enforcing independence should be considered.

etc. as the proxy for the candidates' qualifications neither harms nor benefits 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. refer 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" [16, 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. Due to this conceptual closeness of WAE and MB the literature oftentimes presents "We're All Equal" (WAE) and "What You See Is What You Get" (WYSIWYG) as the two opposing worldviews.

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" [16, 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. This is needed when discussing the morality of independence as independence looks at the DS, which of course relies entirely on the OS. Therefore, we need terms to discuss the causes of difference in OS if we want to morally justify enforcing independence in the DS.

We base our extension on the work of Rawls [32] who differentiates between *realized abilities* and *innate potential* (or, as Rawls writes, "native endowments"). Potential is innate to an individual and determined at birth.[3] This could, for example, be their innate intelligence or predisposition (e.g., extroversion) to develop the traits for being a good sales person.[4] The realized abilities represent how good job candidates actually are at making sales at the time when the company is 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 Potential Space (PS).

Figure 1b 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 Potential Space (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.

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 will be defined below), all groups would have the same realized abilities.
- "We're All Equal in the CS" (WAECS), meaning all groups have the same realized abilities (even though it may look differently when taking measurements). This is the literal interpretation of "we're all equal", which implies that all groups are currently equal in their abilities.

As noted earlier, Friedler et al. formally define WAE as equality in the CS, but leave the option that it may also be interpreted as equality in the PS. We explicitly distinguish the two worldviews as we see it as necessary for discussing independence from a philosophical perspective.

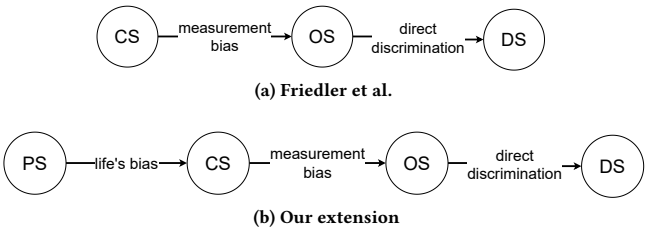
The distinction between PS and CS allows us to define another type of bias. As already stated, Friedler et al. refer to the introduction of group differences from the CS to the OS as "structural bias" while we refer to it as MB since it is introduced through the act of measuring and is dependent on, 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.

3.3 Rules Distilled From the Framework

We will now again consider the framework proposed by Friedler et al. to distill rules about when and why to choose independence as the fairness measure. The extension of Friedler et al.'s framework

[3] "At birth" here should be considered as "at conception" since there is already social influence at the fetal stage [21].

[4] "[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." [32, pp. 56-57]



(a) Friedler et al.

(b) Our extension

Figure 1: Relationship between the spaces and biases.

presented in the previous section will be used in order to clarify the position of the paper.

In order to understand Friedler et al.'s reasons for recommending independence, we have to introduce two key terms that appear in their proposal: Fairness and non-discrimination. We refer to [16] 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.[5]
- *Non-discrimination*: The difference between groups is not (notably) increased from the CS to the DS.

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.

We find that the paper considers 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 separately and derive a proposition summarizing them as a single rule.

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)" [16, 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. This statement should be interpreted to mean that if we assume MB, only independence *can* achieve non-discrimination, but it is *not* a guarantee. In fact, they present an example in which MB is assumed, but there are differences in the CS. In this case, enforcing independence would still be discriminatory. They conclude that enforcing independence only guarantees non-discrimination if WAECS is also assumed.

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

Proposition 3.1. `IF there is MB`[6] `AND WAECS`[7]`, 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" [16, 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.

Further, 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" [16, p. 14] Using independence thus also reflects the assumption that WAECS holds, so `IF independence should be used, THEN WAECS`.

We follow the second part of the rule as follows:

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

Merging the rules. Figure 2 shows both rules and the implications of not following them.[8]

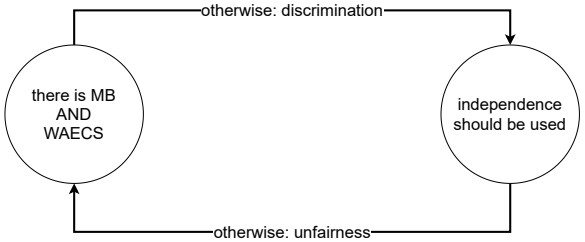


Figure 2: Friedler et al.'s rules for choosing independence.

Note that `IFF x, THEN y` is equivalent to saying `IF x, THEN y AND IF y THEN x`. We can therefore merge Proposition 3.1 and Proposition 3.2 as follows:

Proposition 3.3. `IFF (there is MB AND WAECS), THEN independence should be used.`

To illustrate this rule, let us consider the hiring example introduced in Section 3.1. We first assume that if we split the pool of candidates into demographic groups, then all groups are on average equal in their qualifications, e.g., their ability to sell the hiring company's product. Second, we assume that standardized test scores

[5] This corresponds to Dwork et al.'s *fairness constraint*, which requires that "similar individuals are treated similarly" [12].

[6] What we mean here is "we assume that there is MB". For brevity, however, we will simply write "there is MB" from now on.

[7] Again, we will write "WAECS" to say "we assume that WAECS".

[8] We can find similar interpretations of Friedler et al.'s rules in the literature, e.g. in [6, 26, 39]. While there are subtle differences between their interpretations and ours (which we lack the space to discuss here in detail), we will show in Section 5 that whichever interpretation of Friedler et al.'s rules is chosen still falls short of justifying statistical parity as a general rule.

shown on the candidates' CVs distort their actual qualifications. To illustrate such a case of MB, Friedler et al. recite research, which calls into question the validity of standardized tests to assess verbal aptitude across different racial groups [33]. In this case, independence should be enforced as both WAECS and MB are given.

Note though that the given reasons for enforcing or not enforcing independence ("fairness" and "non-discrimination") are by definition purely mathematical, not moral. However, the terms still imply philosophical meaning. Lipton and Steinhardt [22] refer 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. This raises the question of 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 paper literally, then it only provides mathematical justifications for the introduced rule. However, since the DS includes predictions that are used to make potentially life-altering decisions, it is hard to see how the rule 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 Proposition 3.3: We will view it as providing moral reasons for/against enforcing independence.[9] 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. The first way of dealing with LB in Proposition 3.3 is to preserve the rule and simply apply it the same way when there is LB. This approach does not consider morality and instead prioritizes Friedler et al.'s [16] original reasoning about mathematical properties. These properties are what "fairness" (in its moral sense) is ultimately reduced to when dealing with LB this way. The second way of including LB in the rule is to extend it and make the argument a moral one. For this, we extend the rule 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.

[9] Binns's interpretation in [6] considers this moralized version.

4.1.1 The mathematical extension. Under this extension, we simply apply the rules that we distilled from Friedler et al.'s framework, independently of whether there is LB or not. Proposition 3.3 states that `IFF (there is MB AND WAECS), THEN independence should be used`. In order to see how this rule plays out when there is LB, we consider the case where WAEPS and inequalities in the CS are caused by LB. This means that by hypothesis `NOT WAECS`. It follows from Proposition 3.3 that independence should not be used in this case. The reasoning consistent with Friedler et al. would be: If WAECS is not assumed, enforcing independence violates Friedler et al.'s mathematical fairness property (that individuals close in the CS should be close in the DS) – independently of whether they are equal in the PS or not. 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 moral extension. In [16], the argument in favor of fairness does not provide an explicit moral grounding to define fairness the way it is defined. This definition of fairness is explicitly inspired by Dwork et al.'s [12] notion, which relies on similarity in the CS to determine what makes a decision *fair*. However, it may be objected that similarities in the CS under the influence of injustice do not provide a suitable reference point to define what is a fair prediction or decision [6].

To see this, let us consider the case of credit lending described by Reuben Binns in [6]. We assume that we observe in the OS that women have historically been less likely to repay their loans than men. Binns provides two possible reasons for this. The first is that credit lenders may be more lenient towards men, allowing them delays in their repayment. This can be interpreted as MB: Women and men who are equally "good" at repaying their loans end up with different credit histories (which are proxies for their repayment ability) as men are treated preferentially. A second reason could be that women are more likely to be single parents, which makes the repayment of credits more difficult. This is an example of LB: Due to societal structures, the average woman is less likely to repay her loan.

The question is then whether differences in the OS between men and women (i.e., their credit histories) should be considered to be just. Binns argues that if they are not considered to be just (e.g., because they are unjustly caused by gendered social structures), then enforcing independence should be considered.

We may argue that both of the given reasons for the difference in the payment history (the existence of MB and LB) are unjust. First, we could say that it is *unjust* (and not merely, inefficient) for a decision about the individual to be taken when MB exists. Second, we may argue that it is unjust (and not merely, inefficient) for a decision about the individual to be taken when LB exists. This suggests that there is a common moral *reason* for why decisions influenced by MB or LB are unfair. We will now attempt to identify this common moral cause.

For this, we will follow the philosophical analysis, which is explicitly invoked by Binns. This analysis will explain why both MB and LB cause unfair predictions and decisions. It appeals to the question of *responsibility* which asks whether individuals are responsible for predictions others make about them that impact their well-being. This responsibility may fail to obtain either because the

measurements that decision-makers have (the OS) do not reflect people's choices (i.e., people are not responsible for MB) or because the construct that is measured (the CS) does not reflect people's choices (or both). Thus, one should ask whether the OS or the CS is the result of *choices* or of *circumstances* individuals cannot control.[10] When MB exists, the individuals who are judged in a biased manner do not control the bias 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. We can thus describe the moral view behind proposition Proposition 3.3 in the following way: if WAECS and MB is assumed, then differences in the DS are unjust on responsibility grounds.

Similarly, one may consider people's actual abilities and behaviors as responses to the peculiar circumstances in which people happen to be born and grow up (which are not up to them). Thus, people are not responsible for the influence of those peculiar circumstances, i.e., for their LB. It follows that people should not benefit or get harmed or, more generally, be treated differently because of LB, *which manifests itself in the CS*. Hence, by parity of reasoning – according to the above moral interpretation of Proposition 3.3 – if we seek to eliminate the influence of MB on the decision, we should also seek to eliminate the influence of LB on the decision.

Conclusion of the moral extension. This suggests the following view as the natural extension of Proposition 3.3, i.e., the view that `IFF (there is MB AND WAECS), THEN independence should be used`:

Proposition 4.1. `IFF WAEPS AND (there is MB OR`[11] `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 moral argument for removing the influence of MB was that it was neither morally neutral, nor merely inefficient, but actually *unjust* as 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 (in the metric of what matters ultimately for justice, e.g. well-being) for which individuals are not responsible are unjust [10, 13]. 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 [32].

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 [32]. Yet these views also converge on many real world cases: for example, current US society is arguably very unjust according to both views.

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 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 Just LB at group-level. In the discussion above, we are saying that not all LB is necessarily unjust because some LB might simply be caused by personal choices. However, one may question whether it is possible for personal choices to cause unequal outcomes not only between individuals, 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. The question is, therefore, whether *just* LB is actually possible on a group-level (which is the relevant level when talking about enforcing independence).

[10] In addition, in the case of MB (but not LB), a distinct moral argument can be given as to why a decision affected by MB is unfair, which is based on merit. The conventional view of merit is that it is based on what people do [25, 28], 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 (A has not contributed more to society than B).

[11] When using `OR`, we are referring to the logical operator $\vee$, which means that the statement is true if either one or both operands are true.

In reply, there are at least two ways of showing the possibility of just LB. First, we consider moral views that differ from luck egalitarianism. Consider the view that the influence of parents who read bedtime stories to their children and in this way cause unequal IQ, i.e., differences in the OS, is never unjust [23], a view that contradicts luck egalitarianism [36]. If that view is correct, inequalities that have been created by reading bedtime stories 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.

Second, just LB may exist under a luck egalitarian view if we do not assume WAEPS. Let us consider the case of a genetic disease which is more common in a specific population due to the founder effect. Spending more resources on the detection and treatment of this disease for the population that is most affected (e.g., more medical check-ups, financial support for therapies) is by definition a form of LB. This is because if we assumed WAEPS, investing more in the detection and treatment of this disease for people in the particularly affected population is the sort of circumstance that would create inequalities in the CS, i.e. the rest of the population would on average suffer more from the effects of the disease. However, in a world in which we are not all equal in the PS (as is the case with this genetic disease), the increased spending on the more affected population could plausibly be seen as a case of *just* LB. Under a luck egalitarian view of justice, for example, this LB would clearly be just because it mitigates an inequality in the CS for which individuals are not morally responsible (i.e., that without intervention, one group is more likely to suffer from the disease).[12]

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*. 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. Whether independence should be enforced or not thus depends on one's view of what LB is just.

[12]Understanding this difference between just and unjust LB can help us understand how our extension of Friedler et al.'s framework relates to Mitchell et al.'s [26] proposition to differentiate between two notions of biased data: "statistical bias" and "societal bias". "Statistical bias" occurs between what they refer to as the "world as it is" (i.e., the CS) and the "world according to data" (i.e., the OS). It is thus simply another term for MB. "Societal bias" is introduced from the "world as it should and could be" to the "world as it is". One may think that the "world as it should and could be" is equivalent to our PS and that "societal bias" is thus equivalent to LB. However, the PS is a purely descriptive notion while the "world as it should and could be", while never defined, suggests a normative concept: To define what the world ought to look like, a philosophical concept is needed. Such a philosophical concept might morally require the introduction of just LB that ensures that the "world as it is" reflects the "world as it should and could be". In this case, our reading of Mitchell et al. is that there is no "societal bias" whereas our model would note that there is LB, but that this LB is just. Our extension thus gives us the tools to describe the existence LB without yet making normative judgments about it. Such normative judgments are only required when differentiating just from unjust LB.

4.3 Extended Rules: Final Formulation

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). As we have shown, the theory of justice one adheres to determines one's judgment about the justness of the LB. This in turn 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 3.3 to include LB is not Proposition 4.1, but rather the following more nuanced view:

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

In what follows, we will consider Proposition 4.2 as a natural extension of Proposition 3.3. However, we want to focus our attention on the implications of Proposition 4.2 in the scenario in which MB does not exist in order to simplify our discussion somewhat. We will refer to the remaining 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.[13]

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.`

In the hiring example, WAEBI is given if, for example, (1) at birth, all demographic groups have the same average potential to become hireworthy sales people, but (2) this equality is lost because one group has, on average, less money and is thus more likely to have to work odd jobs instead of doing unpaid internships. The resulting differences in the candidates' qualifications and proxies like their CV are thus considered to be unjust and should be corrected by enforcing independence.

5 TWO COUNTEREXAMPLES AGAINST EXTENDED RULES

We will now examine if Proposition 4.2 holds up as a general rule for when to enforce independence. Clearly, it does not represent a general rule if we can find cases in which this rule does not apply. The goal of this section is to see if we can find such cases.

As stated before, we focus our search for counterexamples on 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`. When we assume that 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

[13]Note that this second assumption logically leads to the assumption of unjust inequalities in the CS.

(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 counterexamples will show 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 WAEBI, THEN independence should be used

It is now time to offer a convincing counterexample to the view that (absent MB) independence should be enforced 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 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, once 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.[14] 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

[14] 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.

(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 [11]. 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., [32]). The objection against the utilitarian is that it does not respect the *separateness of persons* [32]. Our argument against independence does not imply the utilitarian conclusion, so it is is not vulnerable to this objection.[15]

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.

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, which is based on economic and political theories of equality of opportunity and which provides indications for when independence should be used [18]. 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 us 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 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.[16] 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 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

[15] Our argument is a contractualist one, not a utilitarian one [34]. Our thesis that the perfect accuracy predictor is just is so robust because it is independently supported by contractualism and utilitarianism.

[16] 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.

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 the eyes of rescue drones 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 [35]. 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*.[17]

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) and there is Measurement Bias (MB). We have introduced the concept of Life's Bias (LB) as a type of bias, which influences how the potential an individual is born with develops into realized abilities. This bias 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 as demographic groups can be equal not only with respect to their realized abilities but also their potential.

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 is the view that one should enforce independence if (and only if) there is MB or if "We're All Equal But There Is Injustice" (WAEBI). In other words, it seems like independence could be justified when taking on 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: the first clearly showed that unjust LB does not always morally require enforcing independence; the second made it plausible that (even in the absence of MB) injustice is not required for the use of independence to be morally justified.

The relatively simple and morally plausible proposition linking WAEBI and independence we analyzed here is thus not universally true. One may object to the first counterexample, saying that it presents a case where what is being distributed is not (uniformly) beneficial, that is, the treatment would be a net harm for many of the subjects. However, it is true of many cases discussed in the algorithmic fairness debate that what is being distributed is not uniformly beneficial: arguably, even being admitted to a university that is too demanding for one's skills might be harmful and being released on parole may not be beneficial for the parolee who in fact reoffends and re-enters prison with a worse criminal record. One may further object that in the first counterexample, considering efficiency alone would produce a fair outcome. We argue that if there are cases in which the efficient solution is clearly and intuitively considered the fair solution, we need a philosophical theory that can explain why this is in fact the case. Our argument thus reveals that a promising line of research may be built by judging the morality of fairness metrics not only based on the question of what causes differences in predictions, but also based on how these predictions distribute utility.

[17] For examples of *unjust* causes of segregation see [3, 8, 15, 30, 37].

ACKNOWLEDGMENTS

We thank our three anonymous reviewers for their helpful feedback. This work was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF), grant number 187473.

REFERENCES

[1] Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 801–809.

[2] 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).

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

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

[5] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for

detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).
[6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 514–524.
[7] 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.
[8] Camille Zubrinsky Charles. 2003. The dynamics of racial residential segregation. *Annual review of sociology* 29, 1 (2003), 167–207.
[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[10] Gerald Allan Cohen. 1989. On the Currency of Egalitarian Justice. *Ethics* 99, 4 (1989), 906–944.
[11] 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.
[12] 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.
[13] Ronald Dworkin. 1981. What is Equality? Part 2: Equality of Resources. *Philosophy and Public Affairs* 10, 4 (1981), 283–345.
[14] Equal Employment Opportunity Commission, Civil Service Commission, et al. 1978. Uniform guidelines on employee selection procedures. *Federal Register* 166, 43 (1978), 38290–38315.
[15] William H Frey. 1979. Central city white flight: Racial and nonracial causes. *American Sociological Review* (1979), 425–448.
[16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).
[17] Philipp Hacker and Emil Wiedemann. 2017. A continuous framework for fairness. *arXiv preprint arXiv:1712.07924* (2017).
[18] 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.
[19] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
[20] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).
[21] Eszter Kollar and Michele Loi. 2015. Prenatal equality of opportunity. *Journal of Applied Philosophy* 32, 1 (2015), 35–49.
[22] Zachary C Lipton and Jacob Steinhardt. 2018. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341* (2018).
[23] Andrew Mason. 2006. *Levelling the playing field: The idea of equal opportunity and its place in egalitarian thought*. Oxford University Press.
[24] 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).
[25] David Miller. 1979. *Social justice*. OUP Oxford.
[26] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8 (2021), 12.1–12.23.
[27] Thomas Mulligan. 2017. Uncertainty in hiring does not justify affirmative action. *Philosophia* 45, 3 (2017), 1299–1311.
[28] Serena Olsaretti. 2006. Justice, luck, and desert. *The Oxford handbook of political theory* (2006), 436–449.
[29] 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.
[30] Garrett Power. 1983. Apartheid Baltimore style: The residential segregation ordinances of 1910-1913. *Md. L. Rev.* 42 (1983), 289.
[31] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.
[32] John Rawls. 2001. *Justice as fairness: A restatement*. Harvard University Press.
[33] Maria Veronica Santelices and Mark Wilson. 2010. Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review* 80, 1 (2010), 106–134.
[34] Thomas Scanlon. 2000. *What we owe to each other*. Belknap Press.
[35] Thomas C Schelling. 1971. Dynamic models of segregation. *Journal of mathematical sociology* 1, 2 (1971), 143–186.
[36] 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.
[37] Christopher Silver. 1991. The racial origins of zoning: Southern cities from 1910–40. *Planning Perspective* 6, 2 (1991), 189–205.
[38] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
[39] Samuel Yeom and Michael Carl Tschantz. 2021. Avoiding Disparity Amplification under Different Worldviews. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Forthcoming.
[40] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery* 34, 1 (2020), 163–200.