Representative Individuals

The demands of fair machine learning are often expressed in probabilistic terms; yet, most of the systems of concern are deterministic in the sense that whether a given subject will receive a given score on the basis of their traits is, for all intents and purposes, either zero or one. What, then, can justify this probabilistic talk? We argue that it can be justified by attending to morally salient aspects of data-driven decision systems (as opposed to, say, the epistemic limitations of particular agents or the identification of some hidden source of stochasticity) and provide a framework for characterizing fair machine learning in probabilistic terms. Our framework identifies the statistical reference classes used in fairness measures with what John Rawls called representative individuals, hypothetical persons who are representative of social positions. We then address the question of how to determine in a principled way which social positions should be represented. We identify, motivate, and critically evaluate three possible approaches.The first is causal: a group should be represented if it appears in a causal explanation of inequality. The second is psychological: a group should be represented if individuals subjectively identify with it. The third is moralized: a group should be represented if membership to that group alters the moral evaluation of the inequality, independently of the satisfaction of the two previous conditions.

CCS Concepts: • **Social and professional topics → Computing / technology policy**.

Additional Key Words and Phrases: fair machine learning, ethics, political philosophy

ACM Reference Format:

. 2022. Representative Individuals. In . ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In fair machine learning, the demands of fairness are often expressed in probabilistic terms. For instance, Fairness and Machine Learning—an influential guide to fair machine learning—defines the fairness criteria it covers as follows [2]:[1]

> *Independence*[2]: for all groups a, b where R is binary,
>
> $\mathbb{P}\{R = 1 \mid A = a\} = \{R = 1 \mid A = \mathrm{b}\}$, that is
>
> $A \perp R$
>
> *Separation*[3]: for all groups a, b where R and Y are binary,

[1] Where $\mathbb{P}$ means "probability of", A is a protected attribute, Y the target variable, and R the classifier or score.
[2] Informally: sensitive characterics should be statistically independent of one's score. So, for instance, the acceptance rate should be the same for all groups.
[3] Informally: sensitive characterics should be statistically independent of one's score, among those who have the trait that is being predicted. So, for instance, the acceptance rate should be the same for all who are qualified.

Manuscript submitted to ACM

$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$ and

$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$, that is

$R \perp A \mid Y$

Sufficiency[4]: for all groups a, b where R and Y are binary,

$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}Y = 1 \mid R = r, A = b$, that is

$Y \perp A \mid R$

It is common for these sorts of criteria—group fairness criteria, which "attempt[] to define fairness in terms of statistical parity"(Fleischer 2021)—to get glosses that make it sound as though these criteria are about equalizing the individual probabilities of certain outcomes.

For instance, independence might get a gloss like,

> "subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class."[12]

That separation,

> "implies that the probability of an applicant with a[...] good credit score to be correctly assigned a good predicted credit score and the probability of an applicant with a[...] bad credit score to be incorrectly assigned a good predicted credit score should both be the same for male and female applicants." [12]

And of sufficiency that it implies

> "for any given predicted probability[,] [...] the probability of having an actually good credit score should be equal for both male and female applicants."[12]

But the notion that group fairness criteria are about equalizing individual probabilities–which is what the above glosses seem to imply—is difficult to make sense of. Further, it is difficult to understand how achieving these measures is to play any significant role in justifying a system if they aren't about equalizing individual probabilities.

To see this, consider the fact that the overwhelming majority of the algorithms we are concerned with in discussions of fair machine learning are deterministic in the sense that whether one will receive a given score on the basis of their particular set of traits is, for all intents and purposes, either zero or one. While it might be true of a given deterministic

[4]Informally: having the trait that is being predicted (in the running example of the footnotes, being qualified) should be statistically independent of one's protected attributes, given one's score (that is, being accepted).

algorithm that, say,

> Subjects in both protected and unprotected groups are assigned to the positive predicted class at equal rates.

This—which is all independence really guarantees—is not enough to ensure that

> "subjects in both protected and unprotected groups have equal [individual] probability of being assigned to the positive predicted class" [12],

To see this, consider the following case:

> **Unfair Admissions**. A law school will select applicants from a pool that is 50% Black and 50% White, where 10% of the White population is wealthy and 10% of the Black population is wealthy. The institution can only admit 5% of applicants, and knows that about half of those admitted will accept. Further, it knows that wealthy students are more likely to have connections that will get them jobs, can predict wealth and race with perfect accuracy, and knows that having a high placement rate and diverse student body is the best way to improve the reputation of the school. It thus extends admissions to the wealthy 10% of applicants, a group of students that is 50% White, 50% Black, and 100% wealthy.

In such a case, it is true that relative to the protected groups "Black" and "White," independence is met. That is, subjects in both protected and unprotected groups are admitted at equal rates, i.e., 10%. But this does not mean that subjects in both protected and unprotected groups have a 10% probability of being assigned to the positive predicted class: arguably, 10% have a probability of one, and 90% a probability of zero. Call the problem of justifying (or explaining away) talking about machine fairness in terms of equalizing probabilities *the equal probabilities talk problem.*[5]

Now, one might say that the equal probabilities talk problem can be dealt with handily: group fairness measures aren't about individual probabilities, they're about group-level ratios and nothing more. The glosses given above might be representative, but they are representative of fast and loose talk and nothing more.

We do not think that the equal probabilities talk problem can be so easily explained away. On the assumption of a very broad sort of liberalism—where justifications are owed to individuals—it is hard to see how group-level measures

[5]Given that our illustrative example only shows that satisfying independence does not involve evening personal probabilities, one might reasonably wonder whether it is not a problem for separation of sufficiency. Let us here show that it is.

Begin with separation. Consider Unfair Admissions 2: a school is considering White and Black students and trying to predict whether they have advanced math skills, which will allow them to skip a grade. As in Unfair Admissions, 50% of students are White, 50% are Black, 10% are wealthy, and the wealthy are 50/50 Black/White. Space is limited and the school will let those in the 90th percentile skip. Wealthy students can afford private tutors who teach to the test, allowing their tutees to appear competent but without improving their underlying competence. For simplicity assume competence is randomly distributed throughout the class, that one gets a tutor *iff* wealthy, and that one gets into the 90th percentile *iff* tutored. Black and White students have equal false positive rates, but, again, the probability of receiving a false positive is zero among the unwealthy and non-zero for at least some of the wealthy. In other words, the Black/White false positive rate is even at the group level but this does not mean that the odds among or between either of these groups is even.

Let's now turn to sufficiency. Consider Unfair Admissions 3, where everything is the same as before except the school lets more students, allowing a few–and only–truly competent unwealthy students to skip (and, say, in even numbers across groups). Now attend to the fact that sufficiency is met: the ratio of qualified students who were identified as qualified to students identified as qualified will be the same across groups. Yet, the probability of being competent given one's score is not. Among the unwealthy it is one, and for at least some wealthy it is not.

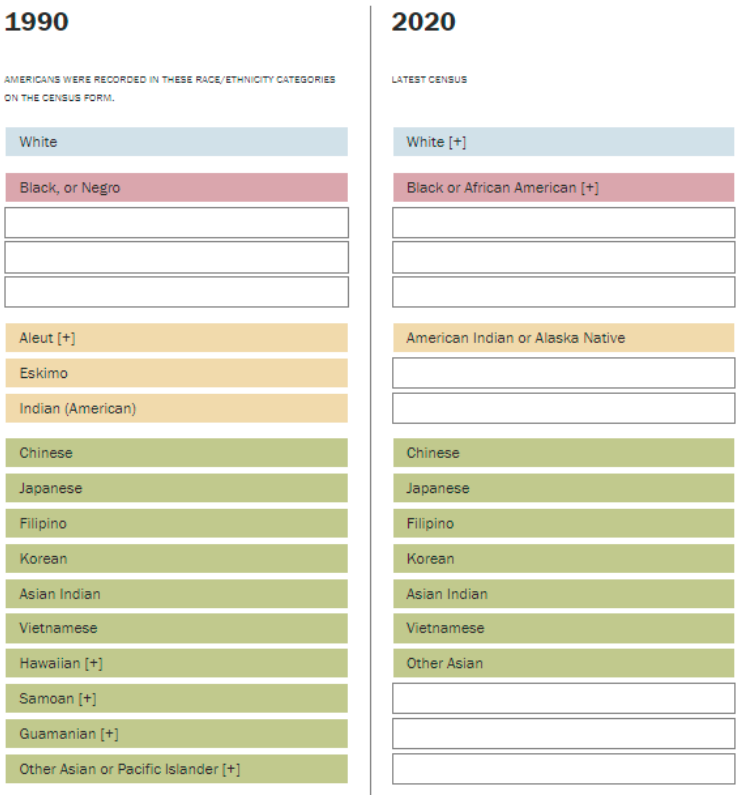

Fig. 1. From the Pew Research Center's "What the Census Calls Us" (https://www.pewresearch.org/interactives/what-census-calls-us/).

could be useful, worth talking about or aiming for, if they are merely about achieving certain group-level ratios. And liberalism in the broad sense just mentioned is, we take it, the dominant paradigm. And yet, group fairness is the dominant approach to fair machine learning [4]. For these reasons, the most obvious way out of the equal probabilities talk problem does not seem very promising.

To complicate matters, any talk of individual probabilities is itself problematic. [3] To determine the probability of an event (or proposition), we must locate it in a class (or, at least, something like a class). But it is often underdetermined which class an event belongs to. For instance, we may want to assess the probability that members of protected groups are falsely identified as recidivists. How do we want to carve up these groups? Should, for instance, our categories for "american indian/alaskan native," "asian," and "pacific islander" more closely match the 1990 or 2020 U.S. census categories?

How we frame our categories will—in a large range of cases—affect our answer to the question of whether a system is fair relative to some fairness measure. For instance, it might be the case that within a system—and relative to all other groups—the false positive rate among Aleuts is very high but among Eskimos very low, such that these differences are erased if they are grouped together but pronounced enough to deem the system unfair if treated as separate groups.

Generally, we can ask: What should the preferred reference class for group level measures be? Call this *the other reference class problem* (to distinguish it from *the* reference class problem[6]).

Returning to our discussion of liberalism, we take it that there is a default presumption in terms of justifying certain decision systems—such as credit rating systems—to the individuals that are subject to them, and that at least part of the justification will involve giving individuals similar chances as certain counterparts of theirs. Given the other reference class problem (as well as the reference class problem), individuals may have similar chances under one description (say, "American Indian or Alaska Native") but not at another (say, Aleut), so the reference class problem—which is independent of the equal probabilities talk problem—has as much bite as the equal probabilities talk problem.

Let us add that there is a special case of this problem that has received, in our view, too little attention in the machine learning literature, which involves the phenomenon of intersectionality and intersects with the other reference class problem. This is the problem of knowing which intersections matter. It can't be the case that we want to require equal chances across all intersections. In the first instance, this is impossible in many cases. More importantly, it does not discriminate between what we might call genuine intersections (e.g., the intersection of "Black" and "Woman") and spurious ones (e.g., "drinks batidos de mamey" and "has read Mother Night"); we think that a satisfactory solution to the other reference class problem will lend guidance on how to answer this question.

2 MORAL JUSTIFICATION, AVERAGE BEHAVIOR, AND INDIVIDUAL DECISIONS

When reality is described probabilistically, it is often for practical reasons. Perhaps one could, in principle, give a deterministic description of the phenomenon they have chosen to describe in probabilistic terms. However, providing such a description might be impractical and a probabilistic approach just as good, given one's aims.

For example, a factory may produce a certain number of substandard articles. When considering that number, we might choose to describe the phenomena at a level that is appropriate for our purposes and treat it as a stochastic phenomenon, even if it is not. This is because our purposes are ultimately practical: we might be deciding a margin when determining the minimum price, at what level to invest, what insurance rate to assign, and so on. For such decisions to be made well, it is often sufficient to approximate general tendencies at a certain level of abstraction and treat the phenomenon as though it is probabilistic.

This thought generalizes in at least two ways. One is that one might accept a certain probabilistic treatment of a phenomenon that is less accurate than some other more accurate probabilistic treatment for purely practical reasons. For instance, we might choose a less accurate probabilistic assessment of the factories we insure than a more accurate probabilistic assessment due to considerations of cost. The other is that we might let practical factors decide the level of abstraction at which we describe the phenomenon. For instance, we might abstract away entirely from talking about what kind of products are made (shirts, tires) and just talk in terms of units or, perhaps, risk of monetary loss (abstracting away from talk of units all together).

With this in mind, our guiding thought is as follows. The algorithms that are discussed in the machine learning literature are justified (when they are) within certain margins of accuracy and at certain levels of abstraction that are set by certain goals of the system that are achieved when predictions are made at a certain scale.

[6]We take the reference class problem to be everyone's problem, in the sense that it isn't just a problem for frequentists. Following Hájek, we think "the classical, logical, propensity and subjectivist interpretations [of probability] also fall prey to their own variants of the reference class problem."[5] We likewise take the other reference class problem to just be everyone's problem.

Consider COMPAS, a software that predicts recidivism and is used by U.S. courts in decisions such as whether to grant parole. Let us ask whether any of the features having a weight in the COMPAS prediction offers in and by itself a moral justification of a possible decision to grant parole. Examples of such features are:

- Current charges of the defendant (e.g., homicide vs. burglary).
- Whether the person was on probation at the time of offense.
- Whether the person is suspected to be an admitted gang member, the number of prior juvenile felony arrests, the number of specific prior offense arrests (distinguishing between different types of offense, e.g., felony property, murder, assault, family violence, sex offence, drug trafficking, drug possession etc.), the number of violations of parole.
- Whether the inmate was raised by both parents, the mother, the father or was adopted, whether the relatives have been arrested, whether friends were arrested or used drugs, the number of times the inmate has contact with his family, whether the inmate was ever suspended or expelled from school or repeated a grade, how often the inmate gets bored, several indicators of personality traits, moral attitudes such as whether the inmate agrees with the statement that a hungry person has a right to steal.

Some of these features (e.g., current charges) might indicate that the defendant was more responsible for their crime. Others, however, might indicate diminished responsibility (e.g., whether the inmate was raised by both parents). But they do not factor into COMPAS in this way, they are used strictly for the purposes of predicting future behavior. Thus, the only possible moral justification for a tool like COMPAS resides in the socially beneficial effects it could produce if its predictions were accurate enough at a certain scale. Crucially, the moral justification involves a characterization of the model or algorithm in terms of its consequences, abstractly described.

3 AT WHAT LEVEL OF ABSTRACTION SHOULD THE FAIRNESS OF A SYSTEM BE DESCRIBED?

In this section, we explain the relevance of a certain way of considering individuals and groups that has been elaborated within liberal political philosophy for the practical task of selecting statistical reference classes in a way that is not ad hoc. Before we introduce the ideas from moral and political philosophy necessary for understanding our project, it will be helpful to zoom out and give a general description of our goal.

Our goal here is, first, to explain the style of answer that we are proposing in this paper. We do this in the service of achieving our main goal, i.e., making sense of the widespread use of probabilistic measures referring to large classes of individuals as measures of the fairness of algorithmic systems that are (in many cases) deterministic. In order to do this, we take consequentialism in moral theory as a reference point. We then highlight a version of rule consequentialism that is peculiar in its choice of units of analysis. While rule consequentialism in standard formulations is *individualistic* (it considers the impact of the rules on the sum of individual utilities), the theory we have in mind selects *groups*, not *individuals* as units of analysis. The version of consequentialism we're sketching here is, in our view, a possible interpretation of Rawls's approach in *Theory of Justice*. While we support this interpretation with references to Rawls's work, we are happy to admit that it is only one possible interpretation and other readers may diverge in their interpretation of the important Rawlsian concepts that have inspired our account. Our purpose here is not to provide a piece of Rawlsian exegesis, but to use some ideas of Rawls as a way to justify referring to groups without ascribing intrinsic moral importance to them. This allows us to locate our theory within the boundaries of, broadly speaking, a *liberal* political view.

We'll begin with an observation: even if one is not consequentialist about the conception of justice or fairness used for the sake of evaluating algorithms, a meaningful debate of these properties seems to take some kind of generally consequentialist justification of their use as a precondition of this debate being morally worthwhile.

In the debate about the ethical qualities of algorithms, you very naturally end up discussing rules in terms of the (total, or average) consequences to which they lead. Mentioning the average performance of an algorithm and explaining how this contributes to a valuable goal of some human agents (while harming no one, or producing less harm than the benefit produced) contributes to providing (at least part) of a consequentialist justification for it. We are not advancing the rather implausible idea that the fairness of an algorithm can only be assessed in rule consequentialist terms (e.g., an algorithm is fair if and only if it maximizes the utility produced for the aggregate of society). Ours is, rather, the more plausible contention that any understanding the fairness of a model/algorithm must focus on the rule, and involve a description of its average or total outcomes, in a manner reminiscent of the evaluation of rules by rule-utilitarianism.

We are going to use an interpretation of John Rawls's *Theory of Justice* as a model for evaluating rules in terms of their consequences on groups, as opposed to individuals. Rawls may seem an odd reference, in so far as he is a liberal political philosopher, and liberals (including Rawls) think about just institutions as institutions that could be justified to individuals and that take individual rights, first and foremost, seriously. Yet, Rawls's analysis of justice as property of a system of rule, rather than of the outcomes it produces, leads to a particular way of analyzing the relations between individuals and the systems that affect them. Rawls thinks that principles of justice apply first and foremost to fundamental social institutions (the "basic structure" of society). They require (among other things) considering how the least advantaged individuals are affected by them. But unlike some economic measures of inequality (e.g., the Gini index), that are often computed by while ignoring groups in a population, Rawls introduces the idea of a representative individual, that stands for a social position (corresponding to a group of individuals) defined by the institutions in question.

Interestingly, it seems that Rawls's *Theory of Justice* also provides a model for thinking about inequalities between groups without abdicating to the basic liberal tenet that justice involves justification to individuals. As we shall explain in more detail in what follows, injustice is not assessed by determining the wrongness of the individual event that Ann Coleman, born on April 15, 1957, in Birmingham, Alabama, found a job as a waitress at age 32 where she makes $1,200 a month with what she should have been paid, given a moral standard of morally deserved salary. It is assessed by asking whether a large class of individuals (those with earnings less than a half of the median income), by making $1,200 a month, could have ended up with a higher salary if the rules had been different. Evaluating the basic institutions of society requires, according to Rawls, abstracting from the details of the concrete lives of individuals and considering how the rules affect the relevant social positions, also called "representative individuals", abstractly conceived.

According to Rawls, when we evaluate whether social institutions are just, we do not have to consider all the individual consequences that they produce, e.g., whether every individual receives from the system advantages and disadvantages that match the contribution of each individual to the system, or that maximize the sum total of utility. Rawls labels that view of justice "allocative". His own, "procedural" solution requires abstracting from the individual-level assessment of the impact of rules. Instead of asking "how does this specific set of social institutions impact on every single distinct individual in our population?", Rawls requires that one asks how they affect representative social positions (or "individuals").[10] In the standard formulation of the Difference Principle, for example, these are individuals described solely by one feature, namely, the socio-economic class in which they are born and in which they stay for their entire life. The Difference Principle is not satisfied by those institutions that maximize the expectations of the least well-off individual in society, and then the next-to-last well off, etc… If this were what it asked, it could not be

really considered applicable to real-world societies, which are very large, and in which these facts are often not known, and even when knowable, considered of private, not of public interest. The Difference Principle aims to maximize the expectations of the least well-off group of people, who are described by Rawls (given a simplifying assumption of healthy functioning) in terms drawn from sociology or economics, (he mentions two possibilities: unskilled labor or individuals with less than half the median income). It is clear that the "unskilled laborer" and the "individual with less than half the median income" is a statistical abstraction just as the average american.

On one influential interpretation, Rawls realizes the idea of the moral division of labor, namely the idea of transcending the values that are relevant in small-scale interpersonal relations, when deciding on questions of basic justice. [11] One advantage of providing an account of justice that abstracts from this set of values is, according to Scheffler, that of delivering a liberal view, one that can accommodate significant variance and diversity in the "arena of small scale personal relation". [11] It is precisely because it is sufficient for justice that social institutions produce a certain average outcome for representative positions. A political class that aims to achieve justice can pursue an objective - the maximization of the expectation of the least well off representative individual (where this individual individual corresponds to a position in the socio-economic structure defined by basic institutions). This can be achieved without specifying in detail the advantages and disadvantages that each concrete individual ought morally to obtain by virtue of their, often unique, more fine grained characterization.

Our proposal is therefore to adopt the idea of "representative individuals" from Rawls in the explanation of why it makes sense to use an abstract account of the behavior of a multifaceted socio-technical system that, in and of itself, may be described at several different levels of abstraction.

Just as Rawls and Scheffler did, one can describe algorithms in terms of how representative individuals fare through them. Simply put, one considers how selected social positions or representative individuals, e.g., Black women, fare under the algorithmic system, as opposed to asking how that individual (e.g., John Doe) fares, and then the next one, etc. This seems to be accepted implicitly by those who take the satisfaction of any kind of group fairness definition to be relevant for fairness. The philosopher's goal in analyzing such justification should be, at the very beginning, to make such moral presuppositions explicit.

So, for example, for many algorithmic systems, there is a level of abstraction that explains individual predictions both deterministically and narrowly, at the level of each individual involved. In theory, one could assess the fairness of COMPAS by considering every actual decision carried out on the basis of a recommendation of the algorithm; and one could compare the well-being and moral character of each and every individual classified as "high-risk" of recidivism with the well-being and moral character of each and every individual classified as low - risk. But even if privacy allowed it, this is perhaps not the best way to capture the important facts about the fairness of this system.[7] If we treat group fairness metrics and standards as salient, we implicitly accept an understanding of fairness that considers average or total behavior and, to the extent that it considers distributional effects and not only whole-population measures, considers discrete, morally salient, groups that are representative of the social reality in question, rather than individuals. In contrast to rule-utilitarianism, the relevant average behavior is not defined *exclusively* by utility promotion for the sum total of individuals, but the *distributional* implications of the average behavior of the rule are considered as well.

[7] Notice that, the original article by Pro-Publica about COMPAS starts with a piece of story telling that seems to be doing exactly this. The article starts by telling and comparing the personal histories of Vernon Prater and Brisha Bornen, a White man who was classified as low-risk and a Black woman who was classified as high-risk. And places two very large pictures, with the names of Bernard Parker, a Black person who was rated high risk; and Dylan Fugett, a White person who was rated low risk, above the title. Perhaps it is a must of contemporary journalism to provide this kind of story telling. It should be evident that very little can be learned from the examination of individual cases.

4 SPELLING OUT THE HYPOTHESIS: POSSIBLE MORAL REASONS FOR ABSTRACTION IN FAIRNESS DESCRIPTION

In the previous section, we offered something of a programmatic, partial justification for treating group fairness measures as normatively significant from a liberal point of view. That is, we provided the beginning of an answer to the equal probability talk problem.

This does not mean that group measures are firm footing. Far from it. For one, which measures to use is left entirely open here, and it is not a problem we will address. Further, how to use measures is left entirely open as well. We will now address one important aspect of this latter question, the question of which groups to use when applying fairness measures. In other words, we will now turn to the other reference class problem. In giving our answer to this problem, we take that we are also shedding light on the equal probabilities talk problem, given that on our justification–one that works by way of representative individuals–the two problems are intertwined.

We shall provide three hypotheses about the reasons that make certain groups of individuals morally salient. The first hypothesis is that one ought to consider a group as a representative individual, if membership to that group plays a role in the explanation of social inequality. The second hypothesis is that one ought to consider a group, if it provides individuals with social identities with which they subjectively identify. The two accounts are not identical, for one could identify with a group even if group membership is not, or is no longer, involved in an explanation of social inequality or oppression. The third is moralized: a group should be represented if membership to that group alters the moral evaluation of the inequality, independently of the satisfaction of the two previous conditions.

It may be objected that in this section of the paper, we are re-inventing the wheel, since groups are of obvious moral relevance in the philosophical debate on discrimination.[1] We defend ourselves by noting that the debate on discrimination differs from what we are doing here in terms of the underlying hypotheses. In analytical philosophy at least, philosophical theories of discrimination start from definitions that are designed to be compatible (and even imply) platitudes about discrimination, e.g., that it can be often wrong to discriminate, that discrimination concerns groups, etc. While the agreement on any of these assumptions may not be universal, certain claims are regarded as platitudes by a sufficiently large proportion of participant in the discourse. Then, the recognition of certain claims as platitudes is regarded as constitutive of a speaker's mastery of the relevant concepts. If so, they can justifiably act as constraints on the acceptability of the philosophical theory or definition advanced. In this vein, a leading article summarizing the debate in analytical philosophy says that "Discrimination against persons [...] is necessarily oriented toward them based on their membership in a certain type of social group."[1] This is not the case for fairness: few would argue that "unfairness against persons is necessarily oriented towards them based on their membership in a certain type of social groups. Here we are not trying to make sense of group fairness definitions by assuming that they are oblique references to the concept(s) of discrimination. We stick to the literal designator most common in the current debate and assume, for argument's sake, that statistical definitions of (group) fairness try to say something about *fairness*, rather than discrimination. Thus, the fact that they refer to groups is not, for us, a platitude that we must account for, but almost a mystery to be made sense of.

4.1 The Causal Account

The causal account takes a group to matter for the description of "representative individuals" if and only if group membership figures in the best explanation of the unequal distribution of advantages and disadvantages in society. Theories of oppression that identify certain groups (e.g., men, whites) as systematically privileged or advantaged and as

oppressors and other groups (e.g., women, people of color and other minorities) as underprivileged, disadvantaged or as oppressed can fill the epistemic burden. The idea of causal relevance is suggested by Lippert-Rasmussen's definition of group as socially salient "if perceived membership of it is important to the structure of social interactions across a wide range of social contexts".

The causal account provides an interpretation for why some groups, but not others, are considered as representative individuals in ordinary group fairness metrics. In relation to the problem of "which groups or intersections of groups" should one then consider, it provides the following answer: those and only those that figure in the best explanation of the causes of social inequalities. But it does not provide an explanation for all the features that are considered, even in some of simplest, more widely used, group fairness metrics, such as the one requiring equality in the false positive and false negative rates.

One important limit of this approach is that it is not suitable to explain a group fairness definitions such as equalized odds which aims to equalize the frequency of a positive prediction by groups that are identified not only by their socially salient properties (e.g., gender) but also by the observed actual outcome in test data. For example, in the case of parole decisions, separation requires that all actual re-offending defendants (something that can only be assessed when looking how the algorithm fares with historical data) receive a positive prediction with the same frequency independently of their group. "Being an actual reoffender" is not a cause of social inequality in the sense that is in question; it is, more plausibly, an effect. So, it is possible that the causal account is not sufficient to explain the representative individuals that are implicitly taken to be salient by those who discuss algorithmic fairness by asking whether group fairness definitions are satisfied.

Another important limit is that general claims about oppressors and oppressed groups are holistic and, thus, difficult to verify or falsify through specific pieces of evidence. Also partly because of this they sometimes end up being controversial. Moreover, they tend to raise some degree of moral, political, or ideological opposition, for example they may be accused to fuel divisive rhetoric that is a hindrance to fixing the root causes of these problems.

4.2 The Psychological Account

The psychological account takes a group to matter for the description of "representative individuals" if and only if group membership is one that individuals subjectively identify with and, therefore, it matters to the individual whether the people who are members of such groups could be differentially impacted by the use of an algorithm or model. Identification, as we mean here, is not a distinct psychological phenomenon, but a placeholder for different possible phenomena that may play roughly the same role in a moral justification. In other words, different psychological phenomena may be sufficient, but not necessary, for identification. One way in which people identify with groups is by having feelings of solidarity and solidarity-motivated dispositions towards them. For example, a homosexual person cares about the fate of other gay people in society and is disposed to spend part of her income to contribute to campaigns protecting the rights of other homosexual, even if that person does not take herself to be at risk of discrimination due to especially favorable social circumstances.

To a large extent, the psychological and causal accounts converge. But they do not need to. It is logically possible for a group to remain psychologically important because, for example, the group has been oppressed in the past, even if it no longer is. Identification is primarily fundamentally a psychological, and subjective, phenomenon. By saying that identification is a subjective phenomenon one means that, even if the theory that treats one's group membership as the cause of a social disadvantage turns out to be false, or to be undecidable (for lack of sufficiently persuasive evidence, or lack of a sufficiently robust social science explanation, or too many competing hypothesis) the simple fact that people

reasonably perceived themselves as disadvantaged on account of their group membership may be enough for the type of identification that makes the group socially salient and therefore morally salient for fairness.

This account may be criticized because it makes the description of a model/algorithm as fair depend on purely subjective phenomena. But there could be at least two reasons why subjectivity may play such a role. First of all, individuals who have been members of oppressed groups in the past, or groups that are still disadvantaged in some spheres of life but not all, may develop a "reasonable sense of inferior political status" (Hosein 2018). According to relational- respect- and recognition- based theories of social justice, it can be unjust for individuals not to be able to consider themselves as political equals, for reasons for which they are not responsible, such as the institutional arrangements of the society they live in (Hosein 2018). Moreover, it may be extremely difficult to determine whether a given group membership plays a causal role in explaining the inequality that emerges in a specific domain of application of models/algorithms. When evidence is lacking or open to different interpretations, it may not be unreasonable for individuals of those groups who have been historically discriminated or oppressed, or anyway, not put in the position to live as political equals, to be at higher risk of developing a sense of inferior political status. So, there is a moral reason to treat membership to such groups, e.g., women or Black people in America, as special vis-à-vis groups such as people who have read Mother Night.

Second, a justification for considering subjective identification in a moral account of fairness is provided by indirect forms of utiliarianism. Rule utilitarianism may justify taking into consideration inequalities between groups with which people identify, but not between groups that no one perceives as salient. Where act-utilitarianism requires that we make the action that produces the best consequences measures in terms of aggregate human happiness, rule-utilitarianism requires that we choose the rules that produce the best consequences in terms of aggregate human happiness.

Quite simply, people who solidarize with certain groups may suffer when individuals of those groups end up worse off as a result, for example, of the accumulation of errors by algorithms. They may suffer in a special way, that is not comparable for intensity with the way someone who is born on Wednesday may suffer, if other people who are born on Wednesday have worse expectations than people born on Tuesday, because of some weird correlation emerging from the statistical indicators in use by the model. Thus, it makes people generally happier to adopt rules avoiding groups (of equally deserving individuals) being unequally impacted by algorithms, when those groups are those with which individuals subjectively identify. By contrast, adopting rules avoiding groups (of equally deserving individuals) being unequally impacted, due to mere correlations, may not produce any significant utility, when those groups are those with which individuals do not identify, e.g., groups of people who like Disney movies or not. A rule-utilitarian argument could be provided for treating the first type of group membership as "morally salient" in a way that supports the equalizations of prospects (for equally deserving individuals) between the representative individuals of the different groups, where the groups are those with whom people subjectively identify. This may produce more happiness than the alternative, more accurate but less group "fair", algorithm, in the long run. Clearly, the same argument, which considers the actual human feelings of solidarity or humiliation associated with perceived identities, does not justify sacrificing accuracy (and the utility it produces) when feelings of solidarity or humiliation connected to group identities are not affected.

This, however, does not make group fairness reducible to whatever maximizes a specific utilitarian function. Appealing to a rule-utilitarian justification in deciding which groups to consider in a fairness definition is not the same as optimizing the algorithm for maximizing the sum of utility of all individuals. The way in which utility is promoted is, in the former case, less direct and less precise. One is not trying to measure exactly the amount of disutility produced by virtue of inequalities with which individuals subjectively identify and one is not trying to include this measure as a

component of the overall utility goal an algorithm is tasked to promote. The latter approach would be coherent with the moral premises of the rule-utilitarian justification of "representative individuals", but arguably in practice, we will never have sufficiently accurate utility measures to formalize this overall utility goal. However, the apparently "deontological" strategy of selecting psychologically salient groups for special consideration when specifying a fairness goal for algorithms can be regarded as an indirect way to promote a broader notion of social utility than the one to which the fulfillment of the direct goal of the algorithm contributes to.

The psychological account, however, also has limitations. Conditional group fairness definitions partition the world in a way that is not easily explained by the psychological account. They ask whether individuals who are actual defaulters, or actual reoffendants, receive the positive prediction with the same frequency across the different (typically, socially salient) demographic groups. The psychological account seems to apply to the socially salient groups much better than it applies to the identity "being a defaulter" or "being a reoffendant".

4.3 The Moralized Account

It is now necessary to look at a feature of the description of certain models/algorithms that only some group fairness measures, but not all, manifest. The "representative individual" in certain definitions of statistical fairness do not involve a group understood merely as a demographic unity. Consider the group fairness definition of equalized odds, or separation. For a binary classifier, this requires the true positive and true negative rates to be the same between two groups. The true positive rate and true negative rates are not always or generally equalized when the rates of people with a favorable prediction are the same between two demographic groups, e.g., women and men. When the proportion of individuals who are "actual positives" or "actual negatives" differs for women and men, odds can be equalized by models that produce unequal rates of favorable decisions for women and men. In the credit lending example above, equalized odds are obtained when the rate of favorable predictions or decisions is the same between the non-defaulting creditors who are women or men.To be rigorous, the true positive criterion is called "equality of opportunity in [6]. The equalized odds criterion also requires the satisfiaction of the true negative rate, that is to say, defaulting creditors are equally likely to be denied credit. In other words, average prospects of a favorable decision are equalized for the following representative individuals:

> 1) non-defaulting clients who are men 2) non-defaulting clients who are women
> And
> 3) defaulting clients who are men 4) defaulting clients who are women
> But they are not equalized between, for example: 1) non-defaulting clients who are men 4) defaulting clients who are women

It is therefore evident that treating equalized odds as a fairness standard implies a different conception of the "representative individual", from the one that merely describes the relevant individuals by a selected demographic characteristic, in this example, as a man or woman. Such fairness definition presupposes that another feature of an individual is morally salient, namely, being a defaulting client or a non-defaulting client. It is an interesting fact about this definition that it can only be verified by looking at historical data, when we have a significant proportion of cases where the attribute "being a defaulting client" is known. Clearly, equalizing the rates of, say, men and women, deemed creditworthy is not identical to equalizing the rates of non-defaulting men and women deemed creditworthy. To the extent that the

definition is used to make an assessment about the fairness of the algorithm/model in the future, it rests on assumptions about the future being relevantly similar to the past represented in the test data. But apart from this complication, this statistical definition of fairness is not obviously problematic or meaningless. So, it is worth asking what justifies, from the normative point of view, the use of the trait "being an actual defaulter" or similar traits (e.g., "being an actual re-offendant" in the parole case) in general.

The relevance of the attribute "being a defaulting creditor" cannot arguably be easily explained by either the causal or the impact account (as defined). Arguably, defaulting creditors do not differ from non-defauting ones in terms of their role within a functional explanation of inequality or oppression in society, so the causal account does not explain why they contribute to define a relevant "representative individual".

The psychological account also cannot explain it, at least in the form that has been given. For it is not so clear that individuals may develop feelings of identification and solidarity with others who share with them the same destiny of being a loan defaulter (even though this is not strictly speaking impossible). Perhaps some people identify with "being a good creditor" and feel solidarity with that social group. Already the case of the reoffending-non-reoffending distinction appears less plausible: can they really perceive themselves as belonging to different groups in ways that are important to their identity and motivate solidarity? And there may be cases in which there is really not meaningful social identity to be attached to the observed label, relative to which equalized odds are computed, e.g., individuals who actually purchase online courses in STEM subjects, etc. Or anyway, the psychological salience of these groups may not be strong enough to explain the intuitive salience that these fairness definitions seem to enjoy.

There is, however, a possible explanation that shares some significant elements with the impact account. For it may be argued that the relevance of the distinction between defaulting and non-defaulting clients has to do with the differential impact that receiving a positive prediction/decision has for them. The positive prediction is creditworthiness and the consequential favorable decision is, let us suppose, lending money at a favorable rate. It may be argued that while future non-defaulting clients are benefited from the favorable decision, the future defaulting clients are not. Arguably, it would have been better for them to not receive that loan in the first place.

This reasoning is analogous to the psychological account in so far as the defaulting / non-defaulting distinction is morally relevant on account of the difference it makes for the utility generated by the decision. When we assess the impact of algorithmically-generated inequality between two groups, say men and women, we should distinguish between the unequal treatment often and women who are identical in relation to their actual defaulting status and the unequal treatment of women who differ in that respect. For the first inequality may generate a disadvantage for one salient demographic group relative to the other. The other one, by contrast, may be advantageous to all.

Another possible justification of considering the status of an individual qua defaulting client, or, in other examples, re-offending parolee, well-rated employee, etc. could appeal to the idea of merit. For one may argue that actual defaulting clients all equally do not merit receiving a loan and all re-offending parolees equally do not merit being released on parole. Hence, when examining the performance of an algorithm/model with test data, we should distinguish the proportion of those favorable decisions that are deserved (e.g., a non-reoffending parolee receiving parole, a non-defaulting creditor receiving a favorable loan) from the proportion of favorable decisions that are not (e.g., a reoffending parolee receiving parole, a defaulting creditor receiving a favorable loan). The two situations ought not to be mixed, because there is a salient moral difference between them. What is morally relevant, in the merit-based view, is not that being granted a credit, or parole, produces different utility gains for the individuals subjected to the decision, in the case of non-defaulting and non-reoffending ones, compared to the utility produced for defaulting and reoffending ones.

In the merit-based (as opposed to the well-being based) version of the moral account, the fact that these representative individuals differ in what they morally *deserve*. [8]

Our purpose here is not to argue for the above view of desert. We are aware of how problematic such moral judgments are, when critically examined.[9] Our point, rather, is that such moral ideas may explain what makes the equalized odds fairness standard prima facie plausible. If so much is granted, it is certainly interesting to observe that this amounts to conceiving the "representative individuals" as characterized by some morally salient properties, either their capacity to benefit from a decision, or their merit. And this characterization cannot be explained by the causal theory or the impact-theory in the subjective identification version. These, however, appear more plausible when we have to explain the demographic element in the "representative individual" characterization.

Notice that, however, some of the most widely discussed group fairness metrics, such as equality of opportunity (that, for a binary classifier, amounts to equality in the true positive rate), equalized odds (that for a binary classifier, amounts equality in the true and false positive rates), sufficiency (that, for a score on the continuum scale amounts to calibration, and for a binary classifier amounts to predictive value parity, also known as conditional use accuracy equality), are all defined on the basis of both, typically, socially salient characteristics derived from anti-discrimination law *and* characteristics related to the correct prediction in "ground truth" labels, that may be argued to be morally salient. That is, they also distinguish the groups of individuals for whom the predictions are equally successful (or not) and focus on the inequalities that emerge between the intersection of those groups. For example:

	Black		**White**	
	Actually reoffend	Do not reoffend	Actually reoffend	Do not reoffend
Predict: reoffend	70	30	80	20
Predict: not reoffend	30	70	20	80

In this table, the probability of a Black parolee who does not reoffend to be predicted to reoffend and denied parole is 30%, while the probability for a White parolee is 20%. This may feel unjust. Two papers [7, 9] in the literature invite the readers to consult their fairness intuitions, presenting two hypothetical examples of violation of this standard that, quite plausible, for most people do not feel unfair. Thus, this definition of group fairness cannot be plausibly considered a necessary condition of fairness. Still, it is important to notice that the examples used in both involve purely hypothetical groups that are not socially salient. One example [9] considers two groups in which students have been divided for didactic purposes in the course of a class, and the second one [7] considers two rooms, A and B, to which individuals have been randomly assigned, and that have been created purely for the sake of a game of luck. In both cases the groups are also ephemeral: they are constituted in the course of the process and then immediately dissolved, they are assumed to have no causal relevance, and they are not groups with which individuals identify. Our account supports the idea that some justification for the choice of the representative individuals is needed. This is coherent with the prediction that our intuitions about the moral salience of the violation of group fairness may vary depending on the nature of the groups that are considered as representative individuals.

Let us, therefore, bracket the critique provided by those arguments, and concede at least for argument's sake that inequality in the true positive rate feels unfair at least when it concerns some morally and socially salient groups. If that is the case, we notice that all the violations of equalized odds that people care about in practice involve a representative individual that is partly defined by its demographic quality (men and women, race, etc...) and partly by another quality

[8] This seems the moral view underlying the "moral framework" proposed in [8].
[9] As explicitly recognized in [8].

(identified with the observed feature in historical test data) that suggests a level of justification for some inequalities (but not others). This is consistent with all the hypothesis we consider in this paper combined.

5 CONCLUSION

In this paper, we ask why group fairness definitions, widely appealed to in the debate on fairness and machine learning, describe the algorithmic socio-technical systems to which they refer in terms that are probabilistic, where the probabilities have, typically, certain demographic groups as reference classes. We start by problematizing this fact and by arguing that it is, in a certain sense, a practice in need for a justification. We then sketch a general approach that makes sense of the appeal to groups. This approach refers to Rawls's concept of "representative individuals" and shows that assessing fairness in reference to groups may not be incompatible with liberalism. In the second and longest part of our paper we lay forward three distinct rationale that may be offered for regarding various groupings as a representative individuals. These are: a causal rationale, a psychological identification rationale, and a moral justification one. These explanations we offer take the form of sufficient conditions: if the rationale is satisfied, then it is prima facie reasonable to consider the group. Since each rationale is offered as a sufficient, not as a necessary, condition, the three hypothesis are not mutually exclusive. One of them may be *the* reason to consider a group in one case and another one in another; moreover, one group may be considered by appealing to more than a single rationale. Finally, certain combined groupings, e.g., the representative individuals "woman who does not reoffend when released on a parole" and "man who does not reoffend when released on parole" may only be explained by appealing to multiple rationales.

We are aware that we have not provided conclusive arguments in favor of any of these hypothesis. Our aim here is more modest, namely to identify this issue as a plausible research question and sketch a possible way in which it may be discussed philosophically. Hopefully, future work will lead to a philosophical theory that determines which groups should be regarded as relevant and why. We still have not provided a comprehensive moral justification taking certain groups and not others to be relevant, e.g., an account of the necessary and sufficient conditions making it correct, from the moral point of view, to consider a group. And we leave some aspects of our explanation open, for example we do not provide a fully persuasive moral justification for considering groups that are relevant for causal explanations of inequality. It was not our purpose to defend any conclusive view, but rather to frame the terms for a possible debate to be had.

The result of this debate may be practically relevant in guiding the decisions of practitioners with respect to fine grain of the group fairness definitions they use to assess the fairness of these systems.

REFERENCES

[1] Andrew Altman. 2020. Discrimination. In *The Stanford Encyclopedia of Philosophy* (winter 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/discrimination/

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[3] Philip Dawid. 2017. On Individual Risk. *Synthese* 194, 9 (2017), 3445–3474. https://doi.org/10.1007/s11229-015-0953-4

[4] Will Fleisher. 2021. What's Fair about Individual Fairness? 480–490. https://doi.org/10.1145/3461702.3462621

[5] Alan Hájek. 2007. The Reference Class Problem is Your Problem Too. *Synthese* 156, 3 (2007), 563–585. https://doi.org/10.1007/s11229-006-9138-5

[6] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[7] Brian Hedden. 2021. On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs* 49, 2 (2021), 209–231. https://doi.org/10.1111/papa.12189 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/papa.12189.

[8] 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 (FAT* '19)*. ACM, New York, NY, USA, 181–190. https://doi.org/10.1145/3287560.3287584 event-place: Atlanta, GA, USA.

[9] Robert Long. 2020. Fairness in machine learning: against false positive rate equality as a measure of fairness. *arXiv:2007.02890 [cs]* (July 2020). http://arxiv.org/abs/2007.02890 arXiv: 2007.02890.

[10] John Rawls. 2009. *A Theory of Justice*. Harvard University Press. https://books.google.com/books?id=kvpby7HtAe0C

[11] Samuel Scheffler and Véronique Munoz-Dardé. 2005. I—Samuel Scheffler. *Aristotelian Society Supplementary Volume* 79, 1 (2005), 229–253. https://doi.org/10.1111/j.0309-7013.2005.00134.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0309-7013.2005.00134.x

[12] Sahil Verma and Julia Sass Rubin. 2018. Fairness Definitions Explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (2018), 1–7.