AI and Ethics
https://doi.org/10.1007/s43681-025-00675-y

ORIGINAL RESEARCH

The representative individuals approach to fair machine learning

Clinton Castro[1] · **Michele Loi**[2]

Received: 27 August 2024 / Accepted: 4 January 2025

Abstract
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 the statistical reference classes used in fairness measures can be understood as defining the probability that hypothetical persons, who are representative of social roles, will receive certain goods. We call these hypothetical persons "representative individuals." We claim that what we owe to actual, concrete individuals—whose individual chances of receiving the good in the system might be extreme (i.e., either zero or one)—is that their representative individual has an appropriate probability of receiving the good in question. While less immediately intuitive than other approaches, we argue that the representative individual approach has important advantages over other ways of making sense of this probabilistic talk in the context of fair machine learning.

Keywords Fair machine learning · Algorithmic bias · Fairness · Technology ethics · Philosophy of technology

1 Introduction

Consider

> **Hiring Algorithm**.[1] A machine learning system is used to determine which applicants should get a first-round interview. Unfortunately, the machine distributes its errors unevenly: it identifies men who are qualified for the job as worthy of interviews at much higher rates than it does women who are qualified.

It would be very natural to say in this case that the machine is biased in virtue of its differential error rate[2] across groups and that further, in virtue of this, it is unfair.

This thought aligns with how many—including ourselves—are inclined to think about fairness in the context of machine learning. But how, exactly, are certain group-level asymmetries (such as unequal error rates[3]) and unfairness connected to one another when it comes to machine learning?

Here is one intuitive defense recently made explicit by Sune Holm [1]:

> Intuitive defense.[4]
>
> (1) When distributing goods (e.g., callbacks), it is important from the perspective of fairness that

[1] This is a fictionalization based on a real case; see Dastin (2018) for the real-world case.

[2] It is inessential to this example that it focuses on a case where error rates (as opposed to predictive power) is different across groups. This paper does not take a stand as to which of those different—and often incompatible Chouldechova [8]—group-level ratios are preferable.

[3] We are not engaging in questions of which fairness measures apply in which contexts in this paper. For discussion of these issues, see, Hellman [9], Hedden [10], Long [11], Holm [1], Grant [12], and Loi et al. [13].

[4] This articulation of Holm [1] bears some semblance to the articulation given by Castro and Loi [14]. While "the intuitive defense" mirrors Holm—who, we should, mention is channeling Broome [3] in (1)—we take it that the intuitive defense at least roughly represents a fairly widespread and natural thought: Fairness is about giving individuals fair chances, and statistics (such as error rates) give us a glimpse into individual chances.

✉ Clinton Castro
clinton.g.m.castro@gmail.com

[1] University of Wisconsin-Madison, Madison, USA

[2] Politecnico di Milano, Milan, Italy

Published online: 28 February 2025

individuals with similar claims to the good have similar chances of receiving the good.

(2) We can assess whether (1) is met by considering whether the appropriate group-level ratio (e.g., equal error rates) is in proper proportion, as this is what individuals' chances of receiving those goods consists of.

(3) So, we can use group-level ratios (such as equal error rates) to assess the fairness of a system.

Applying the intuitive defense, we can say that in Hiring Algorithm there is unfairness because—as the unequal error rates show—qualified women face lower chances at receiving the good than others with similar claims to the good as them (i.e., qualified men).

In this paper, we will reject the intuitive defense, but this is not the main goal of the paper. The main goal of the paper is to grapple with the difficult problems that the intuitive defense is grappling with and propose an alternative solution to those problems. Our solution is less intuitive on its face, but—as we hope to show—it avoids deep problems associated with the intuitive defense and natural alternatives to it.

The paper is structured as follows.

We first identify two problems[5] with the intuitive defense. One of these problems, *the equal probabilities talk problem*,[6] has to do with how the view understands the relationship between group-level ratios and individual-level probabilities. The other problem, *the narrow reference class problem*, has to do with how the intuitive defense groups individuals into reference classes.

Along the way, we articulate and criticize two natural alternatives to the intuitive defense. One of these, *the subjectivist approach*, understands fairness as about evening subjective probabilities (as opposed to the objective chances of the intuitive defense). The other, *the collectivist approach*, understands machine fairness to be about evening group-level ratios (as opposed to anything like individual-level chances or subjective probabilities). We do not think that either alternative is very promising, but considering them is instructive.

We then introduce our alternative, *the representative individuals approach*. This approach identifies the statistical reference classes used in fairness measures with hypothetical persons who are representative of social roles and understands fairness as demanding that one's representative in the system receive fair treatment.

It is not our goal to argue that our approach is correct; our aim is more modest. We simply aim to show that the representative individual approach evades the problems associated with the intuitive defense, the subjectivist approach, and the collectivist approach; further, we aim to demonstrate that it is plausible and worthy of consideration as an alternative to these views. In the end, we hope to put forth a useful and less problem-ridden foundation, or interpretation, of claims about bias and attendant claims about unfairness in the context of machine learning.

2 Two problems for the intuitive defense

2.1 The equal probabilities talk problem

Let's begin with the equal probabilities talk problem.

The notion that group-level ratios map onto individual chances—which is what (2) (above) assumes—is difficult to make sense of.

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 algorithm that, say,

> Qualified men and women are identified as qualified at similar *rates*.

This—which is all equal error rates implies—does *not* ensure that.

> Qualified men and women have similar chances of being identified as qualified.

To illustrate, consider the following highly stylized case:

> **Alternative Hiring Algorithm**. A machine learning system is used to determine which applicants should get a first-round interview. This machine distributes its errors evenly across genders. It achieves this by always identifying privately educated people as qualified and never identifying publicly educated people as qualified. Privately educated, qualified men and women are distributed across groups in such a way that qualified men and women overall are identified as qualified at similar rates.

In this case, it is true that—relative to the groups "qualified man" and "qualified woman"—error rates are equal. But it

[5] One of these problems has been articulated in Castro and Loi [14]. We articulate the problem here to set up a discussion of how our view does not suffer from this problem.

[6] This problem is also discussed in Castro [15].

should be clear that this does *not* mean that individual qualified subjects in these groups have similar chances of being identified as qualified (some have a 100% chance, and the others zero). Call the problem of justifying (or explaining away) talking about machine fairness in terms of equalizing probabilities (which are interpreted as objective chances in the case of the intuitive defense) *the equal probabilities talk* problem.

In response to this problem, there might be any number of initial reactions.

One reaction is to say that the equal probabilities talk problem can be dealt with handily: individual probabilities needn't be considered at all. What we need to fix are group-level ratios and nothing more.

In response, we do not think that the equal probabilities talk problem can be so easily explained away. This collectivist approach might gain a competitive advantage over the intuitive defense when it comes to talk of ratios and probabilities, but seems to lose something at least as important: compelling grounds for fairness complaints.

Unstructured groups (i.e., mere collections of people)—such as qualified men, qualified women, etc.—*as such* seem to be the wrong locus of concern in a case such as Hiring Algorithm. The *individual* qualified and publicly educated women who faced the algorithm seem to have a distinctive complaint, but this approach seems to give up on this thought. The intuitive defense's talk of individual chances—fraught as it may be—at least pays tribute to the importance of individuals. For these reasons, the collectivist way out of the equal probabilities talk problem does not seem very promising.

Alternatively, one might try to cope with the equal probabilities talk problem by going subjectivist. That is, one might reformulate (1) as follows:

> (1') When distributing goods (e.g., callbacks), it is important from the perspective of fairness that individuals with similar claims to the good have similar *rational subjective probabilities* of receiving the good.

The idea here is that all that agents are owed is that we have every reason to *think* they have similar chances at the good. On such a view, we might justify the use of the group fairness measures by saying that the ratios can be used to define the subjective probabilities we assign to individuals' probability of receiving the good.

This view faces a major difficulty. As Hausman [2] notes, views that treat the subject matter of fairness as subjective probabilities are subject to counterexamples, such as.

> **Hiring Algorithm 2**. The team notices that the original hiring algorithm is biased. They retool the system to remove the bias. Unbeknownst to them, however, the system is rebooted after a power outage and comes back online in its original, biased form.

In such a case, we assign each applicant a rational subjective probability of a callback that is the same. And yet, intuitively, they have not been treated fairly. This is because these subjective probabilities don't actually correspond to objective chances (or whatever the proper surrogate for chances is, given the equal probabilities talk problem).

In response to this challenge, one might put some distance between subjective probabilities and what decision subjects are owed. For instance, one might say that the group-level ratios are *evidence* of a bias, one that should inform our subjective probabilities about the chances individuals face or even of a bias that the group faces.

We indeed think that group-level ratios can play an important evidential role in our thinking about bias. However, we think that this second version of the subjectivist view is challenged, as it gives no account of what fairness is about. The first version, despite its flaws, at least does this: it says that fairness is about achieving certain distributions of subjective chances. In other words, this version does not actually address the problem we are raising. Instead, it simply brackets the problem.

2.2 The narrow reference class problem

Let us now introduce the narrow reference class problem.

Note that in our examples—and this tracks real world practices[7]—the set of features used for the purposes of measuring fairness (e.g., man, woman, qualified, not qualified) are not very rich. That is, they do not give us language for describing individuals outside of a small set of socially salient sensitive characteristics and qualifications. They do not give us language for inquiring about whether, as Alternative Hiring Algorithm demonstrated, individuals are privately educated or not. Put another way: the measures as commonly deployed treat many features as irrelevant.

Whether this is an indictment of the measures as commonly deployed is an interesting question. For now, we will not address that question. Though, we should note here that we think that coarse descriptions of individuals' features are often called for. Further, the representative individuals approach that we will advance can aid in giving a principled defense of this thought, even if it also recommends including features that are typically excluded (such as wealth, which being privately educated would be a proxy for).

Note, now, that the intuitive defense is at odds with current practices and common intuitions about the central

[7] See, e.g., https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

concerns of fair machine learning. According to the intuitive defense, it is important that equal claimants have equal *individual* chances at the good. If we were interested in measuring this, we would need to look at subjects in much more fine-grained ways.[8] At first blush, this might seem like a boon to the intuitive view—it might, for instance, encourage us to use a more fine-grained approach to characterizing individuals. But we do not think that the view benefits from this observation overall. This is because it can't tell us why we typically ignore—as we think we should—factors (typically) irrelevant to questions about the sorts of discrimination that we are concerned with in the context of fair machine learning.

Why is this a problem? Here are two different reasons.

First, if the intuitive view is meant to support current practices but in fact does not, then it fails as a justification of those practices. Even if the practices are not justified, the project of justifying those practices on the basis of the intuitive thought collapses. One project that we find interesting in light of this collapse is whether some other justification can be given that is at least internally coherent. In what follows we propose to do just that.

Second, we take it that, in broad strokes, current approaches to fair machine learning—while imperfect and still in need of further development—are broadly defensible. That is, we take it that, in at least some cases, looking at, say, just race and ethnicity in a fairly coarse-grained way is in fact the proper level of analysis. Zooming in to much finer levels of grain can lead us to lose the forest for the trees, so to speak.

Before moving on, we should touch on how these problems interact with the subjectivist and collectivist approaches. In our view, the subjectivist approach—in adopting (1')—is no different from the intuitive view when it comes to the narrow reference class problem. Both views focus on evening individuals' probabilities and, in virtue of this, demand that we pay close attention to individuals' characteristics to test for fairness.

The collectivist faces their own issue with reference classes. Let us assume that they could say that the coarse-grained reference classes are ones they could justify (perhaps citing that the groups they care about are simply those that are protected by law). Their version of the problem is that their reference classes will, in our view, be too coarse. Let us consider a third case:

> **Hiring Algorithm 3.** The team realizes that the hiring algorithm in Hiring Algorithm 2 has rebooted in its original (bad) form. To compensate, they create a third version of the algorithm that will be extremely biased towards women until the group-level ratios of men and women are even, it will then revert to the second version so that it will be unbiased going forward.

In Hiring Algorithm 3, we think that the women judged by the algorithm in Hiring Algorithm 2 have a fairness-based complaint. This complaint, however, is hard to make sense of on the collectivist approach. At the end, we get the group-level ratio that we were hoping for. Put another way, we think that the right view on fair machine learning will be able to distinguish between Hiring Algorithm 3, where some group of people pass, over time, through three different versions of an algorithm (which is for one period biased against women, another biased in favor of women, and for another unbiased) and

> **Hiring Algorithm 4.** An unbiased algorithm runs for the same period of time and sorts a similar size of people as in Hiring Algorithm 3.

[8] Advocates of multicalibration (e.g., [16]) require looking at whether the prediction is equally likely to turn out to be true for groups that are as small as they can be for a meaningful statistical test (given the data one has). Hébert-Johnson et al. [16] explain their motivation as follows "Calibration is typically applied to large, often disjoint, sets of protected groups, that is, the guarantees are only required to hold on average over a population defined by a small number of sensitive attributes, like race or gender. A stronger definition of fairness would ensure that the predictions on *every* subpopulation would be calibrated, including, for instance, the qualified members from the example above. The problem with such a notion is that it is information-theoretically unattainable from a small sample of labeled examples, as it essentially requires perfect predictions. As such, we need an intermediary definition that balances the desire to protect important subgroups and the information bottleneck that arises when learning from a small sample" [16]. In other words, their view seems to be that only de re individual probabilities would lead to fair decisions, but these are as a matter of principle unattainable with ordinary data-driven methods. Multi-calibration would provide the closest feasible approximation to this ideal. This argument resembles A. J. Ayer's [17] argument according to which the "narrowest class in which the property occurs with an extrapolable frequency" ([17], 202) would exclude causally irrelevant statistical data. However, as argued by Oberdiek [18], there are two fundamental difficulties with this alleged solution. First, there is no uniquely correct narrowest causally relevant reference class, because "beginning with any particular initial reference class is arbitrary and because narrowing the initial reference class in any particular way is arbitrary" ([18], 28). Second, Oberdiek argues, wider reference classes can be more causally relevant than narrower ones. He offers the example of the risk of being killed in a car accident, arguing that the risk of an individual is best captured by the reference class "pedestrian, driver or passenger" than the reference class "driver" even for the individual who is a driver (e.g., has a driving license). In terms of our approach, the relevant point here is not that the risk computed relative to the reference class "drivers" is inferior in *causal-explanatory* terms. Rather, the broader reference class is more relevant from the *moral* point of view, because it better represents the *interests* of the relevant parties (your interest to *avoid a deathly accident* as a driver is *aligned* with your interest as a pedestrian and as a passenger).

The collectivist approach will, we think, struggle with a case like this, because it is indifferent to what might be owed at the individual-level, that is, it is indifferent to the sort of complaint that the women have in Hiring Algorithm 2 if it happens in the context of Hiring Algorithm 3. We understand this problem as an offshoot of our earlier criticism of the collectivist view: in losing its connection to individual-level complaints, it is too indifferent to the sorts of individual-level concerns that drive many of our concerns about fairness.

Now that the two problems that motivate us have been explained, we turn to our positive proposal.

3 The core idea

Before developing the representative individuals approach in detail, it will be helpful to have an overview of it in place. The core idea is this: one way to show that a system is fair is to show that it gives fair treatment to suitable representatives of anyone who might face it, and this strategy can be adapted to fair machine learning such that it:

- Does not run afoul of the equal probabilities talk problem
- Is sensitive to individual-level complaints
- Delivers intuitive results about all of the Hiring Algorithm cases
- Does not run afoul of the narrow reference class problem

On the representative individuals approach, we can understand fair machine learning in fairly familiar probabilistic terms; however, the probabilities under discussion are those faced by *representative individuals*—hypothetical individuals who represent a social role (i.e., a complex of categories that individuals are understood as belonging to). The probabilities that the representatives are assigned, in turn, could (but needn't, as we discuss later) be defined (or at least informed) by the frequencies[9] at which the individuals they represent get the good in question.

What remains to be done, of course, is to fill in the details of this picture so we can make good on the claim that it indeed has the attractive features that we claim that it does.

[9] More specifically, we take it that the relevant frequencies are in fact those that would emerge in a long-run idealized trial, which we might be able to gain an approximate understanding of via inspection of a sufficiently large set of results from the actual use of a system.

4 Fairness and fairness measures on the representative individuals approach

We take it that the ground-level demands of fairness vary from context to context and that this means that the selection of a fairness measure for some context depends on the specific demands of fairness in that context.[10] We take it that the relevant fairness measure for a context will ultimately be explained by reference to what we can call *the relevant normative principles* (which could, but needn't necessarily, be a function of facts about the context as well as fundamental moral ideas).[11] This, as we will soon demonstrate, is important to keep in mind, both in our framework and in general.The relevant normative principles help to both determine the applicable fairness measure for a context and how to interpret it.

While we mean to propose a framework and, thus, mainly want to remain neutral on substantive questions of fairness, we do think that our approach requires one substantive commitment about the nature of fairness, at least when applied to certain systems of rules (such as certain predictive systems). As implied by our discussion above, showing that a system is (un)fair will have two main parts:

I. Determining whether the representatives are treated fairly, and
II. Determining whether the representatives are, indeed, representative.

The relevant normative principles will obviously inform us as to how to go about accomplishing these tasks. But, perhaps less obviously, it is also true that this mode of demonstration will constrain which principles we can even consider. This is because whether our approach is legitimate turns on what fairness is about.

[10] See Castro et al. [19] for a similar observation.

[11] These serve a similar function as the "mid-level egalitarian principles" of Castro et al. [19]. For Castro et al. [19], the proper fairness measure for a context is determined by the proper "mid-level egalitarian principle(s)" for that context (which are, themselves, determined by the fundamental moral principles and relevant empirical facts). While compatible with our view, our view is distinct from this picture. For one, while amenable to egalitarianism, we do not want to commit to the idea that the only type of relevant principle will be an *egalitarian* one. Similarly, we do not want to assert that these are picked out by fundamental moral principles. Generally, our view is much like theirs with regards to the idea that there are normative principles that are contextual which pick out the proper fairness measure for a situation. However, our picture is much more permissive than theirs. We do not specify that the principles must be egalitarian, nor do we have any commitments with regards to the structure of the justification of those principles.

Consider two subtly different views of subject matters of fairness. One, which we might call the *de re* approach,[12] takes it that fairness involves something like directly giving concrete individuals equal individualized chances at receiving some good when they have equal claims to that good (this at least approximates John Broome's [3] approach).[13]

A different approach, which we might call the *de dicto* approach, understands things a bit more indirectly. It's not so much, say, *this applicant* that should receive fair (de re) chances at being interviewed; instead, it is *their role* that should be structured a certain way: namely, that *it* has fair chances attached to it. What individuals deserve on this view is that *their role* receives fair chances.

The representative individuals approach sits much more naturally with (and indeed may require) the de dicto approach. This is because a representative individual's chances can be understood as the chances attached to a role (e.g., qualified woman applicant). It is harder, and perhaps simply incorrect, to justify the representative individuals approach as appropriate on the de re approach. This is because in many cases there is a divergence between the chances that I (described richly, in all of my individuality) and my representative (me, again, but much more thinly described, as, say, a qualified applicant) face.

This naturally raises the question of whether the de dicto approach is, contra the de re approach, the correct approach to fairness. We do not think that it is our place to settle this difficult question here. Instead, we are happy to make two brief, non-definitive, comments about why we think that being wed to the de dicto approach might not be a severe limitation of the representative individuals approach.

One comment is that the two views might not be entirely exclusive. Even if the de dicto approach isn't the *one* correct view, it might still apply to some cases, opening the door to the representative individual approach in at least some situations. It could turn out to be the case that, in some cases de re chances matter but in others de dicto matters. That is, it could turn out that there just isn't *one* fundamentally correct view. Instead, there are two views that cover different jurisdictions, so to speak.

The other comment is mostly sociological, but important: *we* find the de dicto approach to be plausible and believe that we are not unique in thinking that it is. Indeed, we take it that something like the de dicto approach is what undergirds one of the most discussed works in political philosophy of all time, i.e., Rawls' *Theory of Justice* [4] (from which we have borrowed the term 'representative individuals'). This is *not* intended as an argument for the view, but it is an (admittedly speculative) argument against the idea that the conversation cannot proceed until we can argue for the de dicto approach (which we assume is a large task for another paper). Insofar as the de dicto approach is one that many are already happy to entertain (though, perhaps, not under that description), the representative individuals approach seems worth exploring.

5 Determining whether representatives are treated fairly

In order to determine whether representatives are treated fairly, we need to specify a relevant normative principle. Our view does not specify what that principle might be. Again, what we are offering here is a framework that has minimal substantive commitments about the content of these principles. For expository purposes, it will be helpful to discuss a specific relevant normative principle that someone might endorse for reasons extrinsic to our framework. To fix ideas, let us stipulate that the proper principle for the case that we will discuss is.

> **Formal Equality of opportunity (FEO)**, which requires that "[a]pplications are assessed on their merits, and the applicant deemed most qualified according to appropriate criteria is offered the position" [5].

Among other things, this sets the standard for what will amount to (un)fair treatment of a representative individual (and, thus, the individuals they represent). Assuming for the purpose of argument that FEO is true, if we can show that two equally meritorious representative individuals have different chances of being identified as qualified, then one of them has been treated unfairly.

With this in mind, our proposal for fair treatment is this. The chances that a representative has in the system—let us say for this example we are interested in knowing their probability of being identified as qualified ($\hat{Y}$), given the social role they represent (A) and level of qualification (Y) (i.e., $\Pr(\hat{Y} \mid Y \,\&\, A)$)—is approximately equal to the corresponding group-level ratio[14] (i.e., in our example, the number of qualified A's accurately identified as qualified, divided by the total number of qualified A's who applied).

Let us now assume this group level ratio, i.e., the true positive rate, is on this interpretation and in the context of Hiring Algorithm, an appropriate operationalization of

[12] Using "de re" (and, later, de dicto) to conceptualize this distinction is, as far as we know, original to us. These terms, however, are borrowed from established use in philosophy of language. See Nelson [20] for a discussion of how these terms have been used in philosophy of language.

[13] This is, indeed, the approach Holm [1] takes.

[14] Assuming that data on a large and diverse number of cases are available.

FEO.[15] In such a case, we can now offer an underwriting of an intuitive thought that does not run afoul of the equal probabilities talk problem. Namely, that if qualified applicants with certain socially salient characteristics are judged as qualified at a lower rate than qualified applicants from some other group, there is unfairness in the system. This is unfair because it does not put them in roles with equal prospects of success (which, per FEO, they are owed).[16]

6 Determining whether representatives are representative

The representative individuals approach requires us to establish what makes a representative truly representative of those it aims to represent. Assuming that the relevant principle to adhere to in the context is an egalitarian one[17] (such as FEO), the determination has three key components: the qualifications of the individuals being represented, their social roles, and how their interactions with the system should be understood. Each of these components requires careful consideration to ensure our framework captures morally relevant aspects of algorithmic decision-making.

On the question of which socially salient characteristics matter, what we are offering is a framework where the choice of the relevant normative principle (in our running example, FEO) and the reasons underlying that choice fill in these gaps. FEO explicitly states that merit matters. But what about social roles? Here it is helpful to note that FEO's focus on merit is motivated by disdain for group status hierarchies [5]. This offers an insight into how to think about which social roles matter. If our choice of a measure is motivated by a disdain for group status hierarchies, then what we are ultimately after is to disrupt (or, at least, refrain from contributing to) intolerable group status hierarchies. This, in typical contexts, will martial in favor of controlling for the sorts of characteristics we are intuitively concerned with (e.g., race) and against controlling for those typically not warranting concern (e.g., whether one enjoys documentaries).

Our approach, then, offers a principled explanation as to why we can concern ourselves with certain features of individuals and not others. Further, it does this without running headlong into any major difficulties (e.g., the sorts of issues collectivist approaches are saddled with). In the end, the choice of the representative is ultimately dictated by a view of which groups matter (which comes from the relevant normative principles), and different moral and political views can combine with the representative individuals approach to establish different criteria for what matters. As we have seen here, the approach can generate intuitive results while avoiding major difficulties.

We can further shed light on this thought by discussing different methods that could assist in determining which groups matter, keeping in mind, however, that these determinations would ultimately be decided as appropriate or not on the basis of the relevant normative principles.

In deciding which groups matter, we might want to take a causal approach. A *causal account* would take a group membership to matter for the description of "representative individuals" when discrimination against members of the group figures into a plausible explanation of the unequal distribution of advantages and disadvantages in society; that is, when certain sorts of systematic treatment of members of that group contributes to odious status hierarchies. Theories of oppression that identify certain groups as systematically privileged or advantaged and other groups as underprivileged, disadvantaged or as oppressed can will fill in various details regarding which hierarchies are intolerable. Our view, which is a framework, does not have a view on which groups in particular get represented. Instead, it says that this choice is ultimately one that is guided by moral considerations that deem certain inequalities as odious.

There are cases where instead of going causal, we might go psychological. What we could call *psychological accounts* will take a group to matter for the description of "representative individuals" when individuals subjectively identify with a group that is taken to stand somewhere in the pecking order of odious group hierarchies. 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. We include any group taken to be in the hierarchy—including higher status groups—because from the perspective of fairness we might care about both the illegitimate conferring of disadvantages *and* advantages.[18]

[15] We are not assuming that this is generally true. See Castro et al. [19] on this point; we are amenable to the case that they make there.

[16] It is important to remind the reader here that we're not arguing for this view of fairness. This is just an illustration on how to interpret such claims.

[17] This is a safe assumption, as much of fair machine learning assumes some form of egalitarianism. As Reuben Binns notes, "`fairness' as used in the fair machine learning community is best understood as a placeholder term for a variety of normative egalitarian considerations" [21], p. 2,cf. Castro et al. [19]). For a consideration of how to embed non-egalitarian patterns of justice (and some unexpected complications) see Hertweck et al. [22].

[18] The psychological account may be criticized because it makes the description of a model or algorithm as fair dependent on purely

It is crucial to recognize that the causal and psychological accounts identify relevant groups through fundamentally different mechanisms. The causal account identifies groups where disadvantage operates through social structures, regardless of individual awareness—much like how a disease might affect a population independently of their recognition of the condition. In such cases, the group membership figures into explanations of disadvantage through purely structural mechanisms. Even if individuals were unaware of their group membership, the disadvantage would persist through the operation of social structures.

The psychological account, by contrast, identifies groups where disadvantage operates through awareness and identification, analogous to how conditions such as diminished self-esteem might affect outcomes. Here, the individual's recognition of group membership is essential to how the disadvantage manifests; without such awareness, this particular mechanism of disadvantage would not operate.

Let us now turn to the issue of understanding how interactions with the system should be understood. To get a sense of the issue that motivates us, consider Hiring Algorithm 3.[19] In that case, a group of applicants over time pass through a system whose settings change during that period. This case highlights the importance of properly representing individuals' interactions with a system (such as the hiring algorithm). If we approach this case using the broad contours of the representative individuals approach, something that should strike us about this case is that it would be distorting to treat all people as having faced the same system over the period.

Why? One way to think about the issue is to imagine whether we could see as reasonable the following objection to the following claim (taking, for now, as given the idea that 'qualified woman' is the appropriate level of abstraction at which to view the chances of qualified women facing the system[20]).

> **Claim**: The role of 'qualified woman' can be represented with one representative whose features are an amalgamation of all features of qualified women who faced the system over the interval.
>
> **Objection**: That amalgamation, which overlooks the settings—which changed during the interval such that it disadvantaged women at some moments and didn't advantage them as others—doesn't represent *me*, who only faced the system at one of those moments when it was at one of the particular settings. It averages over people who effectively faced different systems!

We would first like to note that we take it that this objection is, on its face, reasonable; thus, we take it that, at least prima facie, we need to have at least three different models of the system, one for each setting during the interval. Note how other accounts, namely collectivist accounts, would have difficulty making sense of this. They do not concern themselves with individuals, so a complaint about a model not modeling an individual's interaction with a system would have no traction.[21] We would also like to note, then, how our approach allows us to very easily ask the right question in this case: can representation involve smoothing over the fact that the system was at different settings when different individuals faced it?

In short, we think that the answer is, "no." In other words, we think that the intuition behind the objection can be substantiated. One way to think about this is that one and the same set of inputs ran through the system at different times would trigger different processes, yielding systematically different results. This certainly makes it seem as though there are, for ethical purposes, different systems at play here. Further, these differences are driven by intentional design choices. Each version of the system that applicants are facing has different goals (e.g., increasing the number of qualified women in the pool of applicants who get callbacks vs. holding that number steady). This suggests that, from a moral point of view, the causal differences initiated by the different settings have moral salience that we must attend to (they are not mere events, such as those that might cause an earthquake to occur at some moment as opposed to another). In effect, the actions being carried out by the algorithm at different times are different actions; they have different goals and the way that applicants are used as means towards satisfying those goals are different when the system is at different settings. Thus, there is good reason to represent this difference in the construction of one's representative individual, which is an attempt to model their interaction with a system.

subjective phenomena. But this, we think, is too quick. Members of groups that have been oppressed in the past may develop a reasonable sense of inferior political status [23]. 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) [23].

[19] Reproduced here for ease of reference: **Hiring Algorithm 3.** The team realizes that the hiring algorithm in Hiring Algorithm 2 has rebooted in its original (bad) form. To compensate, they create a third version of the algorithm that will be extremely biased towards women until the group-level ratios of men and women are even, it will then revert to the second version so that it will be unbiased going forward.

[20] However fantastical this might be. We of course think that, given the history of most societies using such systems, we will have to at least consider race and ethnicity and how those categories intersect with being a woman.

[21] Non-collectivist alternatives to our view will fare better here, but, as discussed in Sect. 2, they have their own problems to contend with.

The other views we have canvassed cannot deal with this issue so easily (if at all). If we assume that the system is deterministic when processing the inputs of a specific individual, de re views have the hurdle of figuring out how to even talk about probabilities or chances in the first place. And, again, collectivist views seem as though they would be indifferent to the reasonable complaint that the system "doesn't represent *me,*" as, on that view, individuals are not what matters.

This example demonstrates how our framework can handle complex questions of representation in practice. The representative individuals approach allows us to identify when aggregation across different system states would be inappropriate, while maintaining our focus on morally relevant features of representation. By carefully considering both the theoretical foundations and practical implications of representation, we can better ensure our fairness measures capture what matters from the perspective of justice.

7 Bringing it all together

With the key elements of our picture in place, let us now take stock and show that it has the desirable features listed at the outset of the paper:

- Does not run afoul of the equal probabilities talk problem
- Is sensitive to individual-level complaints
- Delivers intuitive results about all of the Hiring Algorithm cases
- Does not run afoul of the narrow reference class problem

We hope that our reasons for thinking that the representative individuals approach has the first four features are fairly clear at the present moment.

It does not run afoul of the equal probabilities talk problem because by shifting the probabilistic talk as being about chances attached to roles it does not conflate individual-level probabilities and group-level ratios illegitimately. It's sensitive to individual-level complaints because it gives us a method for delivering justifications to individuals; our handling of Hiring Algorithm 3 in the previous section should highlight the distance we see between our view and its competitors. It does not run afoul of the narrow reference class problem because in making fairness about de dicto (as opposed to de re) chances, it does not (contra the intuitive and subjectivist approaches) commit us to conceiving algorithm subjects as falling under unduly narrow descriptions, nor does it (contra collectivism) commit us to erasing important differences between subjects. Finally, as demonstrated throughout (especially Sects. 5 and 6), it performs well on the Hiring Algorithm cases.

One last, previously unmentioned, feature we would like to discuss is the approach's user-friendliness. We take it that the representative individual approach is user-friendly in two important ways.

First, we take it that the view is user-friendly for practitioners. The approach arguably justifies the use of group-level measures, which are the most tractable for use in actual cases. They simply require access to data about cases and the deployment of very simple analytical tools. Further, it justifies the use of broad descriptions of subjects, which facilitates the analysis of cases of interest.

Second, and perhaps more importantly, we take it that the view facilitates productive communication with affected parties, who might be less familiar with the nuances of machine learning. To determine whether a system is fair, all we need to know is whether it treats representatives fairly and whether those representatives represent the individuals they purport to present. This means that, among other things, if we want to instruct affected parties to determine whether a system is fair to them, all we need to do is to help them train their eyes with respect to these two questions: how am I represented? Does my representative receive fair treatment?

We think that these two points combine in ways that can be generative. There are a host of reasons for thinking that engagement with affected parties is important for fairness. Simply working with affected parties and asking them about questions of representation and fairness might help us answer important questions about representation and fairness. Further, involving affected parties might be an important aspect of building procedurally fair[22] systems [6]. Finally, it might also be an important aspect of making politically legitimate[23] systems, as communicating that you are being fair might be an important aspect of the sort of trustworthiness necessary for political legitimacy [7].

These benefits notwithstanding, there are important limitations of this framework that merit discussion.

Relevant to this discussion, it is important to keep in mind that what we offered here operates at a high level of abstraction. Our framework primarily serves to offer an understanding of what talk of "chances" are about in claims such as, "under this algorithm, women applicants face unfair chances of receiving a callback." This is foundational work in fair machine learning, work that has largely been

[22] Where "procedural fairness" is fairness characterized by a "process or procedure that most can accept as fair to those who are affected by such decisions. That fair process then determines for us what counts as fair outcome" [24], p. 4,cf. [6].

[23] That is, systems that are either actually authoritative (i.e., "normatively legitimate") or perceived as authoritative (i.e., "descriptively legitimate"). This will be important when machine learning is used by public institutions, such as police departments. See Purves and Davis [7] for an illuminating discussion of this matter.

left undone and overlooked. Even in the paper that we have primarily been responding to—Holm [1]—what "chances" are about is not part of the main discussion. Rather, ideas on the matter operate as implicit assumptions in the background of the paper.[24] But it is vital to properly understand what we are talking about if we say that an algorithm is fair. Otherwise, we are operating in the dark. And as we have seen, this can lead us astray: the alternative interpretations of what "chances" are about have implausible commitments. The hiring algorithm examples help to illustrate this.

It is partly because it operates at a high level of abstraction that this framework will not result in, e.g., an algorithm for making fair algorithms. As Sect. 5 illustrates, in order to do this, we would need to specify the relevant normative principles for whichever context the framework is being deployed. But the framework does not—and should not—provide the content of the relevant normative principles that it so crucially relies upon. The framework is not engaged in delivering moral advice about what, e.g., fairness demands in the context of hiring. Instead, it is about how to make sense of claims about chances, which will be vital to understanding whether those demands are met. Indeed, there is an important division of labor between the work that this framework can do for us and the work that the relevant normative principles do. Neither can offer complete guidance on their own in the context of machine learning, but the principles cannot be a part of the framework, as sorting out the demands of fairness is a separate task from the one that we have taken on here. The framework offers an interpretation of probabilistic speech in the context of fair machine learning. The principles frame the goals which frame questions of what probabilistic work needs to be done in order to ensure fairness.

So, while what we offered here does have practical import, this import will mostly occur at a foundational or conceptual level. But this does not make it irrelevant to the important work of making fairer machines. As we hope to have shown, it is vital for the success of that project. That being said, it is only one piece of the puzzle. In addition to understanding what "fair chances" are about, we need to understand, for example, what fairness demands in a given context. While we hope that the tools we have given here offer support in that project, they cannot do it all. And this is as it should be: building fair machines is demanding. If the history of discussions of fairness teaches us anything, it is that this will be a large and multifaceted undertaking. We hope to have contributed one small part to that much larger project.

[24] We should note here that this does not detract from many of the paper's main contributions. We have found this to be an extremely illuminating paper, despite its reliance on what we have argued to be a flawed assumption.

8 Conclusion

We hope to have shown that the intuitive approach to justifying the most popular fairness measures falls short in a variety of ways. We also hope to have shown that some natural alternatives to it—i.e., the subjectivist and collectivist approaches—face serious shortcomings as well. Additionally, we hope to have presented a new and interesting alternative to these approaches that, at the very least, avoids these shortcomings.

Beyond addressing these theoretical shortcomings, the representative individuals approach makes a crucial contribution by providing a principled framework for determining which differences between subjects matter for fairness assessments. Rather than attempting to account for every possible individual difference (as in the de re approach) or ignoring important distinctions (as in collectivist approaches), our framework acknowledges that some differences must be abstracted away while providing theoretical tools to justify which distinctions are important.

This transparent handling of differences stands in stark contrast to existing approaches that either implicitly ignore differences or struggle to justify which ones matter, making it a boon to ground-level work in fair machine learning. For practitioners, this framework provides much needed guidance for selecting and justifying fairness metrics while maintaining connections between abstract fairness principles and concrete implementation decisions. For policymakers, it offers theoretical foundations for existing regulatory approaches while suggesting how they might be refined to better capture the complex nature of algorithmic fairness.

Our framework also opens several promising directions for future research. These include developing more sophisticated methods for defining and updating representative individuals as social conditions evolve, as well as creating practical tools for implementing the framework in different contexts. While significant work remains to be done, we believe the representative individuals approach provides a valuable foundation for advancing both the theory and practice of fair machine learning.

Acknowledgments We are grateful to three anonymous referees for helpful comments on this paper.

Funding Support for open access was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. Michele Loi's contribution to this work was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF), grant number 187473, the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement 898322.

Declarations

Conflict of interest No conflict of interest.

References

1. Holm, S.: The fairness in algorithmic fairness. Res. Publica. (2022). https://doi.org/10.1007/s11158-022-09546-3
2. Hausman, D.M.: How health care can be cost-effective and fair population-level bioethics. New York Oxford University Press, Oxford (2023)
3. Broome, J.: Selecting people randomly. Ethics **95**(1), 38–55 (1984). https://doi.org/10.1086/292596
4. Rawls, J.: A Theory of Justice, 2nd edn. Harvard University Press, Cambridge, MA (1999)
5. Arneson, R. "Equality of Opportunity." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall (2008). http://plato.stanford.edu/archives/fall2008/entries/equal-opportunity/.
6. Wong, P.-H.: Democratizing algorithmic fairness. Philos. Technol. **33**(2), 225–244 (2020). https://doi.org/10.1007/s13347-019-00355-w
7. Purves, D., Davis, J.: Public trust, institutional legitimacy, and the use of algorithms in criminal justice. Public Aff. Q.Aff. Q. **36**(2), 136–162 (2022). https://doi.org/10.5406/21520542.36.2.03
8. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
9. Hellman, D.: Sex, causation, and algorithms: how equal protection prohibits compounding prior injustice. Washington Univ. Law Rev. **98**(2), 481–523 (2020)
10. Hedden, B.: On statistical criteria of algorithmic fairness. Philos Public AffAff. **49**(2), 209–231 (2021). https://doi.org/10.1111/papa.12189
11. Long, R.: Fairness in machine learning: against false positive rate equality as a measure of fairness. J Moral PhilosPhilos **19**(1), 49–78 (2021). https://doi.org/10.1163/17455243-20213439
12. Grant, D.G.: Equalized odds is a requirement of algorithmic fairness. Synthese **201**(3), 101 (2023). https://doi.org/10.1007/s11229-023-04054-0
13. Loi, M., Herlitz, A., Heidari, H.: Fair equality of chances for prediction-based decisions. Econ. Philos. **40**(3), 557–580 (2023). https://doi.org/10.1017/S0266267123000342
14. Castro, C., Loi, M.: The fair chances in algorithmic fairness: a response to Holm. Res. Publica. (2022). https://doi.org/10.1007/s11158-022-09570-3
15. Castro, C.: Broomean(ish) algorithmic fairness?. J. Appl. Philos. (2025). https://doi.org/10.1111/japp.12778
16. Hebert-Johnson, U., Michael K., Omer R., and Guy R. "Multicalibration: Calibration for the (Computationally-Identifiable) Masses." In *Proceedings of the 35th International Conference on Machine Learning*, 1939–48. PMLR. https://proceedings.mlr.press/v80/hebert-johnson18a.html (2018).
17. Ayer, A. J. "Two Notes on Probability." In *The Concept of a Person: And Other Essays*, edited by A. J. Ayer, 188–208. London: Macmillan Education UK (1963) https://doi.org/10.1007/978-1-349-01903-8_7.
18. Oberdiek, J.: Imposing Risk: A Normative Framework Oxford Legal Philosophy. Oxford University Press, Oxford, New York (2017)
19. Castro, C., O'Brien, D., Schwan, B.: Egalitarian machine learning. Res. Publica. (2022). https://doi.org/10.1007/s11158-022-09561-4
20. Nelson, Michael, "Propositional Attitude Reports", The Stanford Encyclopedia of Philosophy (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2024/entries/prop-attitude-reports/>.
21. Binns, R, Fairness in machine learning: lessons from political philosophy (2017). Conference on fairness, accountability, and. transparency, New York, Proceedings of Machine Learning Research, Vol. 81, p. 1–11.
22. Hertweck, C., Heitz, C., & Loi, M. (2024). What's distributive justice got to do with it? rethinking algorithmic fairness from a perspective of approximate justice. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (Vol. 7, pp. 597–608).
23. Hosein, A.O.: Racial profiling and a reasonable sense of inferior political status. J Polit PhilosPolit Philos **26**(3), e1-20 (2018). https://doi.org/10.1111/jopp.12162
24. Daniels, N., Sabin, J.: Setting limits fairly: Learning to share resources for health, 2nd edn. Oxford University Press, New York (2008)