Algorithmic Fairness and Resentment

Boris Babic (University of Toronto)
Zoë Johnson King (Harvard University)*

May 12, 2023

Abstract

In this paper we develop a general theory of algorithmic fairness. Drawing on Johnson King and Babic's work on moral encroachment, on Gary Becker's work on labor market discrimination, and on Strawson's idea of resentment and indignation as responses to violations of the demand for goodwill toward oneself and others, we locate attitudes to fairness in an agent's utility function. In particular, we first argue that fairness is a matter of a decision-maker's relative concern for the plight of people from different groups, rather than of the outcomes produced for different groups. We then show how an agent's preferences, including in particular their attitudes to error, give rise to their decision thresholds. Tying these points together, we argue that the agent's relative degrees of concern for different groups manifest in a difference in decision thresholds applied to these groups.

1 Introduction

The artificial intelligence and machine learning (AI/ML) market is presently valued at over 50 billion dollars (Rimol, 2021), with applications in almost every industry from finance and banking (Veloso et al., 2021) to logistics (Gordon, 2021) marketing (Hall, 2019), medicine (Benjamens et al., 2020; Babic et al., 2021b) and criminal justice (Chohlas-Wood, 2020). As more and more decisions with major impacts on individuals' lives become wholly or partially automated, the potential moral and legal ramifications of this automation are becoming an increasingly salient public policy issue.[1]

*Both authors contributed equally to the writing of this paper at all stages of development.

[1]See Babic et al. (2021c) for an overview of AI/ML risks and Lander and Nelson (2021), for a recent call for fundamental legislative reform on regulating AI/ML from White House advisors on science affairs. Some of that reform is already underway. The European Union's General Data Protection Regulation (2016) purports to provide a right to have automated decision

Scholars and policymakers are concerned that we may be inadvertently designing AI/ML devices in such a way as to unwittingly produce or reinforce structural group inequalities. This is especially relevant when classification decisions disproportionately impact historically disadvantaged groups. For example, an AI/ML college admissions system may admit disproportionately less students from historically marginalized groups;[2] an AI/ML lending system may be more likely to refuse to lend to minority groups than non-minority groups, or it may lend to them at disproportionately disadvantageous rates;[3] and an AI/ML system used in bail decisions may predict that defendants from minority racial or ethnic groups are more likely to recidivate than white defendants (Angwin et al., 2016). Our collective unease reaches its peak when disparities like these are produced by AI/ML systems that are so-called "black boxes", meaning that the estimated function relating inputs to outputs is not typically comprehensible at an ordinary human level (Babic et al., 2021a).

The purpose of this paper is to begin a conversation about an alternative way of thinking about algorithmic fairness to that found in the existing literature. We articulate a theory of algorithmic fairness that is (A) grounded philosophically in Strawson (1982)'s "quality of will" approach to moral responsibility, (B) motivated economically as an instance of Becker (1957)'s taste-base discrimination, and (C) articulated mathematically in the language of Bayesian decision theory. The resulting approach is a threshold theory of fairness – a simple approach, with a rich interpretation. It is the interpretation of the threshold approach, and its theoretical underpinnings, that we see as the main contribution of our paper.

To begin, consider a non-algorithmic example. Adela must decide whether to lend Evan $10. Being a rational Bayesian, she will lend him the money if the posterior probability that he repays her is sufficiently high. Suppose Adela construes sufficient height as a probability of 80% or greater. This, as we will later explain, ordinarily implies that Adela takes the cost of lending Evan money when he doesn't repay to be four times as bad as the cost of refusing to lend him money when he would in fact have repaid. Now suppose that John asks

making not based on "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade-union membership" (Council Regulation 2016/679, 2016 O.J. (L 119) 51) although the scope of such a right is a matter of debate (Gerke et al., 2020). Meanwhile, Canada's Bill C-27 would require an organization to provide an explanation of the prediction, recommendation or decision produced by its algorithmic system, including the "principal reasons or factors that led to the prediction." Bill C-27, 44th Parliament, 1st Sess. (Canada 2022).

[2]See e.g. Schwartz (2019). This kind of case corresponds to the now famous example of St. George's Hospital Medical School in London in the 1980s. While such glaring discrimination – taking points off for non-Caucasian names – is less likely with modern algorithms, giving less weight to certain groups in more subtle ways still occurs, as we will later discuss in much more detail.

[3]For example, it has been documented that graduates of historically black colleges, such as Howard University, may be getting on average disadvantageous interest rates on their loans (Arnold, 2016).

Adela for $10 as well, and suppose that her threshold for lending to John is 50%. This implies that Adela regards the cost of lending John money when he doesn't repay as exactly as bad as the cost of failing to lend him money when he would repay. The central idea behind our approach is that this means that Adela *cares more* about the harm done to John if he is mistakenly denied a loan than she does about the harm done to Evan if he is mistakenly denied a loan. In particular, she is treating John's costs on par with her own, whereas Evan's costs are discounted fourfold. As we will explain, it is these attitudes toward Evan and John that give rise to the associated thresholds applied to their requests to borrow money. Thus we can use the thresholds to determine that Adela cares more about John's welfare than she does about Evan's. As a result, we say, she is to that extent treating Evan unfairly.

This is a fundamentally different way of connecting false positives and false negatives to fairness than that in the prevailing approaches in the literature. We do not define unfairness in terms of observed disparities in outcomes, since, on our view, these are neither necessary nor sufficient for unfairness (as we will explain shortly). Rather, we encourage *using* outcomes as evidence to draw inferences about the decision-maker's attitudes, with the latter being the basis for conclusions about equal treatment or the lack thereof.

The paper proceeds as follows. In Section 2, we briefly discuss by way of background the extant literature on algorithmic fairness in general, and we motivate what, we think, ought to be the focus of any theory of group discrimination: the decision-maker's attitudes. We do this by drawing on the quality of will tradition in moral philosophy. In section 3, we develop more formally a general theory of fairness – and a model for measuring the extent of unfairness – that can appropriately capture the philosophical approach previously summarized. In section 4, we compare the way we employ thresholds to the way thresholds appear in some of the existing literature. We also discuss in more detail the relationship between observed outcomes and how they bear on the decision maker's attitudes.

2 Bias, Fairness, and Attitudes

Given the attention algorithmic bias has received, one would expect a lively discussion of what it could possibly mean for an artificial intelligence/machine learning (AI/ML) system to harbor a bias. Prior to discussions of how to detect or measure algorithmic bias, one would expect a theoretical discussion of what the bias of an algorithm could *amount to*, metaphysically speaking. And one would hope to see a discussion that goes beyond definitions based entirely on observed disparities in common summary statistics. That is because these outcome disparities are not what we use to assess accusations of bias leveled at human decision-makers, and thus not what we ordinarily take bias to consist in.

For example, if one were to claim that a law firm is biased against women, one

could not simply point out that the proportion of women employed by the law firm is disproportionately low and rest one's case. There would be many other questions: Are women insufficiently recruited, or are women's careers stymied along their way to partnership? Are there particular individuals whose aim is to stifle women's development? Is there a culture at the firm that makes it harder for women to succeed? And so forth. In short, our understanding of human bias is ordinarily a matter of decision-makers' having certain attitudes: prejudicial attitudes toward individuals or groups, or inclinations to inappropriately favor the interests of certain individuals or groups over others. While it is true that we often look for disparities in observable outcomes as *evidence* of bias, these outcome disparities do not themselves *constitute* bias, and are relevant only insofar as they bear on the decision-maker's attitudes.

Indeed, in the United States, jurisprudence surrounding the Fourteenth Amendment's Equal Protection Clause has focused on whether a discriminatory purpose can be inferred on the basis of the decision-maker's actions. This is made especially clear in *Village of Arlington Heights v. Metropolitan Housing Development Corp.*, 429 U.S. 252 (1977), where Justice Powell, writing for the Court, states that "Proof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause." In the now famous footnote 21 of that case, Justice Powell explains that the plaintiff must prove the state had a discriminatory intent or purpose as its "motivating factor" in the state action being challenged. Absent compelling evidence of discriminatory purpose, the Fourteenth Amendment's Equal Protection Clause is unlikely to be understood to have been violated.

With respect to gender discrimination, the Supreme Court has made clear that a disparate impact is relevant *only* insofar as it sheds light on discriminatory intent or purpose. For example, in *Personnel Administrator of Massachusetts v. Feeney*, 442 U.S. 256 (1979), the Court states:

> When a statute gender-neutral on its face is challenged on the ground that its effects upon women are disproportionably adverse, a twofold inquiry is appropriate. The first question is whether the statutory classification is indeed neutral in the sense that it is not gender-based. If the classification itself, covert or overt, is not based upon gender, the second question is whether the adverse effect *reflects* invidious gender-based discrimination." (emphasis added).

The Court went on to find that in this case, "appellee has simply failed to demonstrate that the law in any way reflects a purpose to discriminate on the basis of sex."[4]

[4]It is worth noting that in this case Justices Thurgood Marshall and William Brennan wrote a dissenting opinion arguing that the degree and foreseeability of a disparate impact can be more probative of discriminatory intent than the Court recognizes. But here even the

These repeated references to intent and purpose turn the law's attention away from the outcomes that a decision-maker produces and toward their underlying attitudes – as is to be expected, given the role that the decision-maker's intention plays in many legal wrongs.

But algorithms do not have attitudes. Or, at least, they do not appear to have them in the familiar way that human decision-makers do, understood as a kind of mental content. (Algorithms do not appear to have mental content in general.) Hence attributing bias to decision-making algorithms seems at least somewhat metaphorical and perhaps unduly anthropomorphic.[5] There is a loose and metaphorical sense of the term 'bias' in which we can speak of the bias of a coin; this sort of speech does not attribute anything remotely like discriminatory purpose to the coin, and is instead understood to be a kind of shorthand for speaking about the coin's physical properties and the impact these properties have on its measures of centrality (for example, changing its center of gravity can 'bias' the average value of a set of tosses). One might worry that talk of algorithmic bias can be no more than an anthropomorphic metaphor of the same sort as talk of the coin's bias. And this worry might be exacerbated by the fact that, despite the substantial and rapidly growing literature on bias and fairness in AI/ML systems, there are surprisingly few discussions of how to make sense of the very idea of an algorithm's harboring a prejudice or its showing special concern for the interests of certain individuals or groups. These ideas, which must be rendered intelligible if anything resembling human bias is to be attributed to a decision-making algorithm, have gone largely undiscussed.

Instead, conversation has centered around the question of which formal conditions a decision-making algorithm would have to meet in order for it to be describable as fair. Perhaps the most well-known of these measures of fairness are *calibration* and *predictive disparity*. An algorithm is calibrated across groups if group members to whom the algorithm assigns the same probability of possessing a certain trait do actually turn out to possess the trait at the same rates. And an algorithm exhibits predictive disparity when it produces false positives, false negatives, true positives or true negatives at a different comparative rate for different groups – for example, when its rate of false positives differs across

dissenting Justices acknowledge that discriminatory attitudes are the primary ingredient, so to speak, and that while impact may be relevant it is so insofar as it indicates these underlying attitudes (with the disagreement focusing on what kinds of facts can provide evidence of discriminatory attitudes – in the Strawsonian terms that we will introduce shortly, the Justices disagree about the relative merits of evidence of *malice* and evidence of *indifference*. We discuss these issues further in Section 3.1.

[5]It should be noted that even if an attribution of attitudes is metaphorical, this does not mean that it is not sensible or useful. Indeed, the legal notion of corporate personhood requires that we impute a similar sort of fictitious mental content to corporate entities; as Chief Justice Marshall states in a well-known case before the US Supreme Court, "The great object of an incorporation is to bestow the character and properties of individuality on a collective and changing body." *Providence Bank v. Billings*, 29 U.S. 514 (1830). Still, it would be good to have a clear idea of what exactly the metaphor amounts to. We do not currently have such a thing.

groups.

For illustration, consider a hypothetical algorithm developed for use by an admissions office in a law school, which uses applicants' GPA and LSAT scores to predict the probability that an applicant would successfully pass the bar exam upon graduation if they were to be admitted (the goal, somewhat simplistically, being to admit the applicants who are most likely to pass the bar). Suppose we observe that of all female applicants that the algorithm deemed 80% likely to pass the bar, 80% of them did in fact go on to pass. And suppose the same is true of all the male applicants predicted to pass with 80% probability. Now suppose that this parity also holds for all of the applicants that the algorithm deemed 70% probable, 60% probable, and so on for each pass probability that the algorithm produces. In that case, the algorithm is *calibrated* across biological sex – conditional on an applicant's pass probability, as assigned by the algorithm, biological sex is irrelevant to whether the applicant will in fact pass the bar.

By contrast, suppose that the algorithm recommends admission for all applicants that it deems over 75% likely to pass the bar, and suppose that we compare all of the female and male applicants admitted on this basis and discover that a far higher proportion of the females than the males did actually go on to pass. In that case, the algorithm exhibits *predictive disparity*: it produces proportionally more false positives for males than for females.

A now-well-known impossibility result (Kleinberg and Mullainathan, 2016) shows that a decision-making algorithm cannot both be calibrated and display predictive parity when the base rates of the relevant trait genuinely do vary across social groups. Much of the recent explosion of literature around algorithmic bias centers around this result. Since each measure accords with a somewhat intuitive idea of what fairness is, the fact that we cannot have both has given rise to the idea that the nature of algorithmic fairness is up for debate (which we think is true) and that it is to be debated largely by comparing the merits and demerits of these and other formal constraints (which we think is false). Some scholars argue that we should aim for calibration at the expense of predictive disparity (e.g. Flores et al., 2016) or *vice versa* (e.g. Hellman, 2020). Others offer further putative conditions on the fairness of decision-making algorithms (see Verma and Rubin, 2018, for an exhaustive overview of proposals).[6]

[6]The above is not intended as an exhaustive survey of extant approaches to thinking about algorithmic fairness. For example, Nabi and Shpitser (2018) and Nabi et al. (2019) develop a causal approach to algorithmic fairness: on their view, bias can be thought of in terms of the presence of an effect of a sensitive feature like race on the prediction along certain causal pathways. Meanwhile, we develop a theory that is grounded in Bayesian rational choice. For someone who wants to view bias from the perspective of economic rationality, and who who is critical of causal learning models – due, for example, to the difficulty in estimating independence and identifying causal models, or to our aforementioned point that causal models reflect "bias" in the sense in which a coin is biased but not in the sense in which a human decision-maker is biased – we have an alternative. However, for someone who is a proponent of causal inference and wishes to capture fairness from this perspective, Nabi and Shpitser (2018) and Nabi et al. (2019) have an answer. As to the bigger question of when and whether

Defenses of these putative conditions tend to proceed by demonstrating their computational tractability and by showing that, when an accordingly constrained model is fit to a particular dataset, it tends to produce results that align with a somewhat intuitive pre-theoretical way of thinking about what fairness is (as compared to a similar model fit without the putative fairness constraint). This has produced a "counterexample game" in the computer science literature of the sort to which philosophers are well-accustomed: subsequent authors can fit the same model with the same fairness condition to a different dataset and show that the results are intuitively unfair, following which further authors can tweak the constraint to avoid the counterexample, and so on (again, see Verma and Rubin, 2018, for many examples).

This is not a good way to approach – nor even to conceptualize – the issue of algorithmic bias. The approach ignores the foundational conceptual question of what it could mean for an algorithm to harbor bias, given that it does not have attitudes in the usual sense that ties attitudes to intentionality and agency in humans. Moreover, it pushes us to evaluate putative metrics of fairness from a quantitative point of view, focusing on the metrics' mathematical properties. But this is not a good starting point. If a concept of bias is unpersuasive from a philosophical point of view then there is little value in showing that it is mathematically tractable or computationally efficient. Instead, we should begin the analysis with moral philosophy. We should first consider what it is to show equal regard for different groups and then consider how an algorithm could succeed or fail in doing so. Only once we have a philosophically defensible theory of the nature of algorithmic bias does it make sense to evaluate it from a mathematical or computational perspective.

Indeed, from a philosophical point of view, both calibration and predictive disparity are manifestly unpersuasive measures of fairness. To begin with calibration: this metric is consistent with an algorithm's assigning proportionally far more of the privileged group members that it classifies into its success category than marginalized group members, and as such it is consistent with a decision-maker's engaging in active, intentional discrimination. Calibration requires only that *conditional* on one's risk score, sensitive characteristics such as race are not relevant. It it is compatible with sensitive characteristics' nonetheless affecting the risk score itself (because they are correlated with or causally relevant to other features on the basis of which risk is calculated). This makes calibration consistent with, for example, the historical practice of "redlining" (see Corbett-Davies and Goel, 2018).[7] Redlining was a lending practice whereby financial institutions are widely understood to have discriminated against racial and ethnic minorities by basing their lending decisions in large part on applicants' geographic locations. Under such a system, the average probability that a

we should use causal inference models for predictive inference, that is beyond the scope of this project.

[7]For a more complete history of the practice of redlining and its insidious effects, see Winling and Michney (2021).

white borrower will secure a loan can be far higher than the average probability that a minority borrower will secure a loan, *even though the algorithm remains calibrated*, simply because the few white borrowers who live in poor neighborhoods default on their loans at similar rates to their minority neighbors – and yet they constitute only a very small proportion of total white borrowers. In this case, conditional on neighbourhood, applicants are denied loans at similar rates across race and ethnicity. But, given the segregation that exists in large North American cities, that is poor evidence of fairness on the part of lenders. For example, it has been shown that in Chicago the mortgage rejection rate for African Americans is three times as high as it is for white applicants. This is compatible with there being no difference in rejection rates at the neighbourhood level (see e.g., Foldessy, 1992).[8] Since these well-calibrated loan decisions can be (and probably were) produced by lenders engaging in active, intentional discrimination motivated by explicit racial animus, calibration is a poor indicator of the absence of bias.

Measures based on predictive disparity do not fare much better. These approaches seek to determine whether a decision-maker is biased by looking at the outcomes that the decision-maker brings about for members of different groups. But the aforementioned legal literature underscores the point that disparate impact is at best *evidence* of bias, rather than itself *constituting* bias. In other words, the outcomes that a decision-maker brings about can play an epistemological role in our attempt to determine whether she is biased (since they can be evidence of bias, assuming that disparate impact is more likely with underlying bias than without), but bringing about different outcomes for different groups is not *itself* what bias consists in.[9]

This point can be underscored philosophically. A familiar lesson from the literature on moral outcome luck is that, whenever an outcome is not entirely within a decision-maker's control, she can bring about different outcomes for different individuals without in fact preferring one to the other — and, likewise, can bring about the same outcomes for different individuals despite being deeply concerned for the interests of one but harboring total indifference or intense hostility toward the other. For illustration, consider the famous example of a driver who accidentally injures a child that, entirely unpredictably, dashes

[8] This phenomenon is an instance of Simpson's paradox, which describes situations where a target property that exists at a group level may not exist to the same extent when the group is subdivided on the basis of a feature that is not independent of the target property.

[9] Someone might argue that we *ought* to identify bias with disparate impact on the grounds that the latter is readily measurable and so the theoretical identification would be useful for courts. However, if this is the *sole* reason offered for identification, then the case for identification is extremely weak. One might just as well argue that we ought to identify bias with sandwiches since it is very easy to tell when a sandwich is present. We hold that epistemology should not drive metaphysics in this fashion; at a minimum, for this kind of theoretical identification to be plausible one would have to argue that the hard-to-observe phenomena and the easy-to-observe phenomena are *relevantly similar*, such that the identification has some plausibility at the conceptual level. And now metaphysics returns to the scene. We hold that the identification has little conceptual plausibility, for reasons described below.

out into the street in front of their car (Nagel, 1976; Williams, 1981). The unfortunate driver does not exhibit *bias* against the child that they injure, nor in favor of other children. It would be a misunderstanding to say that they *display a preference* for all of the children that they do not hit and therefore that they treat the children unfairly. On the contrary, the driver may hold all children in exactly equal regard and may be averse to hitting them to an exactly equal degree. That is why the fact that they happen to hit this particular child (rather than another child, or no child at all) is an instance of bad moral luck. But this simple observation drives a wedge between predictive disparity and a decision-maker's bias: if we want to know whether a decision-maker holds individuals or groups in equal regard, what matters is not the outcomes that she happens to *bring* about but the outcomes that she *cares* about, and the degree to which she cares about each of them.

This is our starting point. It will be the main difference between our way of conceptualizing fairness and bias and the prevailing accounts of fairness in the literature. We frame issues of bias and fairness in a way that makes the decision-maker's attitudes explicit, rather than simply taking disparate outcomes to constitute unfairness. This approach will enable us to connect this contemporary topic to traditional issues in Bayesian decision theory, taking a philosophically defensible account of bias and rendering it computationally tractable. And, in doing so, we will answer the hard question of how to impute attitudes to a classification algorithm in a way that is not unduly anthropomorphic.

We are inspired by the "quality of will" tradition in philosophical and legal theorizing about moral responsibility. This is the approach, traced to P. F. Strawson's pioneering paper "Freedom and Resentment" (Strawson, 1982), that holds that moral agents' reactions to one another are primarily reactions to the *regard* or *care* or *concern* (we use these terms interchangeably) that they display toward ourselves and the people and things that we value. The core of the anthropomorphism worry about algorithms is that it is difficult to see these decision-makers as *agents* given that they simply match input to output following patterns in the data on which they were trained, getting pushed around by the world's causal forces just like a biased coin. And, historically, much the same worry has also arisen for human decision-makers; philosophers have worried that the truth of causal determinism would strip humans of the kind of agency that we need in order to be properly held responsible for what we do. But Strawson's paper made possible a new kind of conversation about agency and responsibility. Famously, Strawson observes that one "central commonplace" of human life is "the very great importance that we attach to the attitudes and intentions toward us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions" (Strawson, 1982, p.62).[10] We care deeply about how others regard us and expect to be treated with *good will* – that is to say,

[10]Strawson's paper is dense and has been subject to much exegetical work. For an influential interpretation, to which we are congenial, see Watson (2004), especially p.221.

with respect for our projects and concern for our interests. As Strawson puts it, "the reactive attitudes... are essentially reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern" (Strawson, 1982, p.70). It is our being disposed to display these reactive attitudes toward someone that constitutes our taking them to be a responsible agent and treating them as such.

Two distinctions within this basic framework are important for our purposes. First, we must distinguish the reactive attitudes that respond to ill will or indifference toward *ourselves* from those that respond to ill will or indifference toward *others*. Strawson draws this distinction, using resentment as an example of the former and indignation as an example of the latter (Strawson, 1982, p.70). Indignation, he says, reflects "the demand for the manifestation of a reasonable degree of goodwill or regard, on the part of others, not simply towards oneself, but towards all those on whose behalf moral indignation may be felt" (Strawson, 1982, p.71). Though Strawson does not explicitly discuss indignation felt on behalf of groups (rather than individuals), we see no basis for a principled restriction here, and so we assume that either form of indignation is possible. (Indeed, in our own case, introspection suggests that both forms of indignation can and do occur.)

That was the first distinction. Second, it is important to distinguish our reactions to agents' *absolute* degree of concern for ourselves and others from our reactions to agents' *relative* degrees of concern for two or more things.[11] Decision-makers can harbor a high absolute degree of concern for each of two people without having *equal* concern for them both: they can care a lot about one person and care even more about the other. This difference in the decision-maker's relative degrees of concern will be manifested in their behavior. It is therefore something to which we can react. And, indeed, we do experience reactive attitudes in response to others' relative degrees of concern; we expect and demand that others treat individuals, groups, and causes differentially only if they actually do differ in their moral importance, showing equal regard for things that are equally important. This is what our interest in fairness and bias amounts to. When someone acts in a manner that displays unequal regard for two (or more) individuals or groups that we think should have been treated equally, we react with indignation. We – the authors – hold that it is this unequal regard that bias consists in. And we hold that indignation at bias or approval of fairness are the reactive attitudes whose presence expresses a demand for equal regard.

To emphasize: as Strawson says, "these attitudes of disapprobation and indignation are precisely the correlates of the moral demand in the case where the demand is felt to be disregarded. The making of the demand *is* the proneness to such attitudes" (Strawson, 1982, p.77). To whom, then, do we make demands

[11] For more on the difference between absolute and relative evaluation of agents' degrees of concern, see Johnson King (2020).

for equal regard? Answer: to every agent such that we are "prone" to experience disapprobation and indignation when we judge that they have displayed unequal regard. And this includes decision-making algorithms. We do in fact experience negative reactive attitudes when we take a decision-making algorithm to exhibit bias. For instance, the general public responded with widespread outrage when ProPublica reported that the COMPAS algorithm was "biased against blacks".[12] On a Strawsonian approach, this demonstrates that we do indeed make the basic moral demand of these algorithms, since making the demand of someone or something just *is* being prone to experience the reactive attitudes toward displays of ill will or indifference from them or it.

Indeed, the widespread outrage that we saw in response to the COMPAS algorithm is to be expected given the emerging empirical literature on people's attitudes toward algorithmic decision-making systems. A number of recent studies have found that algorithmic decision making systems are "recognized as blameworthy agents" and that people are comfortable blaming them for the harm they cause, although the particular dynamic of how that blame is distributed can very quite a bit depending on the type of algorithm, the role of the human, and the context of application (see Lima et al., 2021, 2023; Furlough et al., 2021; Hidalgo et al., 2021).

Of course, we do not respond to algorithms in all of the ways characteristic of holding human decision-makers responsible, since adopting some of these responses to algorithms would be a category mistake. One cannot fine an algorithm or give it a treat, since algorithms have neither money nor mouths and are in any case unresponsive to this sort of Pavlovian conditioning. Nor can we ask an algorithm for the reasons underlying its decisions and expect it to answer – or not, at least, in a way that we can understand. Nonetheless the core of Strawsonian responsibility, wherein we are prone to experience the reactive attitudes in response to a decision-maker's apparent disregard, or unequal regard, for the interests of individuals or groups to whose cause we are committed, remains.[13]

At this point the reader may be skeptical about whether there even *are* such things as the reasons underlying an algorithm's decisions and the algorithm's degrees of concern for individuals and groups. We appreciate these concerns but think that they are misguided. To begin with the former: accounts of reasons-responsiveness in human agents already have to face the fact that humans frequently respond to reasons without representing them *as* reasons and without prior explicit and conscious deliberation. As Arpaly (2000) has memorably put it, "if we were to deny that people act for reasons whenever their actions are

[12] See, e.g., Angwin et al. (2016), Liptak (2017) and Yong (2018).

[13] The literature on human agency and responsibility has long distinguished between varieties of responsibility in a way that underpins the point we are making here (see Shoemaker (2011), Shoemaker (2015); and cf. Watson (1996)). Using the language of this literature, we would put the point by saying that algorithms are neither *accountable* nor *answerable* for their decisions, but those decisions are nonetheless taken to be *attributable* to them.

not the result of deliberation, then we would find that it is uncomfortably rare for people to act for reasons" (p. 506). Arpaly points out that much of what we do and believe is a response to reasons that we do not know are our reasons, sometimes because we are self-deceived but also sometimes simply because we do not think about it (*ibid.*, pp.506-507). For example, perceptual beliefs – I see a cat on the mat and believe that there is a cat on the mat – are rarely, if ever, preceded by conscious deliberation about one's reasons *qua* reasons. Likewise, in cases of fast action – like a tennis player's return of a serve – we can respond to reasons that we do not entertain as such and that we represent and process only on a subpersonal level. Indeed, humans can even respond to reasons in a way that is opaque to ourselves, as when a conclusion finally dawns on someone after a period of gradual accumulation of subtle evidence for it that the agent's attempts at explicit reasoning had consistently overlooked. But all of this means that the fact that algorithms issue judgments (1) quickly, (2) without explicit deliberation, and (3) in response to conglomerations of evidence that they are unable to articulate, should not impugn their status as reasons-responsive since the very same factors are also present for a great deal of human action and belief. It is a desideratum of contemporary accounts of reasons-responsiveness that they include human decisions that display features (1–3), provided that these decisions are still "a good practical conclusion from [the agent's] beliefs and desires" (*ibid.*, p.12).

Those accounts that do so will then accommodate the reasons-responsiveness of decision-making algorithms. For algorithms do treat the statistically significant predictors of a property as reasons for predicting the presence/absence of that property, or its probability, in the ordinary sense of 'reason' proposed by Scanlon (1998) according to which a reason is "a consideration that counts in favour of" an action or attitude.[14] If we want to understand why an algorithm predicted a certain outcome, then we will evaluate its predictors and the way they are combined. For a sufficiently simple algorithm – a linear modeal, say – the predictors will combine additively, with the weights given in terms of their coefficients, so that we can say, for example, "this prediction is based 1/2 on X_1, 1/4 on X_2, and 1/4 on X_3." And we can also make claims to the effect that, *if* the value of X_1 had instead been such-and-such, *then* the prediction would become so-and-so. But this does not mean that an algorithm is just like a coin, helplessly pushed around by the world's causal forces. On the contrary, decision-making algorithms reach good practical conclusions from what are the analogues of their beliefs and desires. This point takes us to the core of our proposal and the main project of this paper: an understanding of the degrees of concern that can be manifested in an algorithm's decisions. We hold that, as with human agents, decision-making algorithms can be seen as fair if the interests of different individuals and groups matter equally to them and biased if certain groups or individuals' interests matter more than others. Hence, the

[14] Indeed, the revised Canadian bill on data protection, cited in footnote 1, requires certain algorithmic decisions to be accompanied by the "reasons or principal factors" that led to the decision.

outcomes produced by the algorithm do indeed play an evidential rather than a metaphysical role. They are evidence of *what matters to it*, where it is the latter to which we respond with the Strawsonian reactive attitudes.[15]

To develop a Strawsonian approach to the issue of algorithmic bias, then, we must investigate what it means for a decision-making algorithm to display care or regard – to an equal degree (fairness), or an unequal degree (bias) – toward individuals or groups. We undertake this task in the next section.

3 A model of bias: the utility threshold

We develop our approach within a Bayesian decision-theoretic framework, according to which a rational agent chooses the act that maximizes expected utility with respect to their probability distribution and utility function and processes information by updating their beliefs via Bayes' Rule. There are two main components of this framework: the agent's beliefs – their probability function – and their desires or values – their utility function. These components are both attitudes, one *doxastic* and the other *prudential* or *moral*. This makes them both the sort of thing that could be a home to the decision-maker's bias. We will ultimately take an approach on which bias and the lack thereof are principally found within the decision-maker's desires or values, and as such encoded in her utility function, although we think that these desires and values can also affect the decision-maker's probability function.

In the microeconomic literature, there are two paradigmatic ways of thinking about bias. The first is doxastic, grounding discrimination in the agent's beliefs (Phelps (1972); Arrow (1972a); Arrow (1972b); Arrow (1974)), whereas the second is conative, grounding discrimination in the agent's desires and values – or "tastes", as economists call it (Becker, 1957). In this section, we first explain these approaches and then situate our own proposal in relation to them.

Our proposal is ultimately quite simple. We will argue, following Becker, that discriminating attitudes are found in the agent's utility function, and we will explain that a rational Bayesian – one who is making decisions by maximiz-

[15]The idea that we respond to agents' decisions as evidence of what matters to the agent is also very familiar from the literature on moral responsibility in humans. For example, Shoemaker writes that "if something matters to me, it is just obvious that I regard it as having some sort of evaluative significance... These attitudes... reflect on me, on my deep self, and in particular on who I am as an agent in the world" (Shoemaker (2011), p.611). And Smith writes that even involuntary responses can "provide an important indication of a person's underlying moral commitments, of who he is, morally speaking" (Smith (2005), pp.241-42). Smith emphasizes that these underlying moral commitments are "not necessarily consciously-held propositional beliefs, but rather tendencies to regard certain things as having evaluative significance ... They comprise the things we care about or regard as important or significant" (Smith, 2005, p.251). This is exactly how we propose to think about algorithmic bias: algorithms do not have anything like consciously-held propositional beliefs about the relative moral status of individuals or groups, but they certainly do have *tendencies to regard certain things as having evaluative significance*, as we will see shortly.

ing expected utility – can be modeled as choosing on the basis of a probability threshold. We describe Becker's approach more carefully below (Section 3.2), but the key point is that, for Becker, the decision-maker's utility is a function of two arguments, the first being monetary gain, and the second being an attitude toward the group of the person that the decision is about. For a discriminating decision-maker this would be a prejudicial attitude or a distaste for a person from that group.

For example, if the decision is binary (e.g. imprison a person or do not imprison a person) and there are two possible states of the world (e.g. the person is in fact guilty, the person is not in fact guilty) and the costs and benefits are all equal, then we should imprison the person if the probability that they are guilty exceeds 50%. If the costs and benefits are not equal then that threshold will change. But, importantly, as long as the decision-maker's goal is to maximize expected utility we can express their decision rule in terms of imprisoning the person if and only if the probability that they are guilty exceeds a certain probability threshold. The important part for this project is that the threshold is expressed as a function of the relative costs of error – i.e., the utility function – and it is these error costs that we interpret in Strawsonian terms.

This relationship between expected utility and probability thresholds is easy to prove (as we illustrate on pg. 24, and pg. 27) and is already well-known. It has been studied in the context of point estimation and hypothesis testing. For example, in Robert (2007), Proposition 2.5.5. establishes this relationship in the general case for estimating a single continuous quantity (namely, the so-called "Bayes estimator" is a fractile determined by the costs of error) while proposition 2.5.7 establishes it in the simpler context of comparing two hypothesis with equal costs of error (in this case, we should choose the hypothesis with the higher probability). Robert suggests that Laplace (1786) was the first to work out a simplified version of this relationship in a study of the proportion of female births in Paris and London.

In recent literature, this result has been applied in Cheng (2013) (Equation 2) (to develop a resolution of the gatecrasher paradox) and Kleinberg et al. (2018) (Theorem 1) (to articulate a trade-off between group welfare and a particular kind of fairness). It is also assumed in Simoiu et al. (2017) (to articulate a threshold theory of fairness in police searches).

Indeed, Simoiu et al. (2017) explicitly adopt the threshold test that we develop in the present paper. We wholeheartedly endorse this sort of approach (as we will discuss in more detail in the final section). The difference between our contribution and their contribution is that Simoiu et al. (2017) assert without argument that different thresholds are evidence of discrimination and they go on to develop a hierarchical Bayesian model for estimating thresholds on the basis of outcome data, where the threshold is treated as a latent parameter and given

a prior distribution. Their project therefore starts where ours ends – i.e., from the assumption that different thresholds imply discrimination. Indeed, they are transparent about this. They state:

> We interpret lower search thresholds for one group relative to another as evidence of discrimination. For example, if we were to find black drivers face a lower search threshold than white drivers, we would say blacks are being discriminated against. pgs. 4-5.

Meanwhile, we take up the prior question of *why* evidence of a difference in thresholds is evidence of discrimination. To use the authors' example: Why does it follow from the fact that black drivers face a lower search threshold that they are being discriminated against? This may seem intuitive, but it still requires a theory of discrimination that connects thresholds to discriminatory attitudes. In what follows we develop such a theory, explaining how different moral attitudes can translate into differences in thresholds.

While our project and Simoiu's project are complementary, then, our contribution ends where theirs starts. We develop a normative account of why thresholds are evidence of discrimination, while they develop an applied model for estimating thresholds from data on outcomes.

3.1 Discriminating beliefs

Phelps and Arrow develop what has come to be known as the *statistical theory of discrimination*.[16] On the Phelps/Arrow model, discriminatory treatment across groups originates in the decision-maker's differing beliefs about members of different groups, where these differences arise as a part of a rational approach to a problem of processing noisy information (Spence, 1973, 1974). For illustration, return to our law school admissions example, and let us now make it somewhat more realistic by saying that the goal of the admissions process is not simply to admit all applicants with a certain chance of passing the bar but, more broadly, to admit the 200 *best* applicants. "Best" is an ambiguous term in this context – does it mean the applicants with the highest LSAT? The applicants who will get the highest grades if admitted? The applicants who will contribute most constructively to classroom discussion? Etc. We might say what we really want is to admit the 200 applicants with the highest *aptitude* to succeed in law, understanding aptitude as an unobserved, latent property of the applicants that explains all of the above phenomena and more. So construed, aptitude is of course a somewhat nebulous concept. But it is no more so than any other unobserved quantity that forms the basis of decisions determining people's fates in life; IQ, EQ, creditworthiness, stress, and recidivism risk, for example, are all

[16] See Phelps (1972), Arrow (1972a), Arrow (1972b), and Arrow (1974). See also Aigner and Cain (1977) and Autor (2003) on whom we principally draw in developing the statistical theory of discrimination below.

estimated on the basis of their observable correlates and then used as the basis for consequential decisions.

Let $\theta > 0$ represent an individual's aptitude. The Phelps/Arrow model proceeds as follows: when students apply to this law school, they submit their application file. This file contains information about θ. But the information is imperfect; some low-aptitude students can learn to "game" the process and come across as much stronger than they are, while some high-aptitude students may not be astute at communicating their true ability. So, let $\hat{\theta} > 0$ represent the noisy information conveyed about θ by the student's application file.

Suppose that all applicants fall into one of two groups of interest, A and B (these could be, for example, racial, gender, ethnic, or religious groups). And suppose that, over the years, the law school's admissions office has seen a sufficient number of applicants from each group and learned that while both groups are approximately normally distributed, group A members are on average higher-performing: $\overline{\theta}_a > \overline{\theta}_b$. Suppose further that the observed signal $\hat{\theta}$ is *noisy*, in the sense that its value can be modeled as coming from a normal distribution with some fixed but unknown variance.

In this situation, if we compare two law school applications that look equally good on paper (i.e., for fixed $\hat{\theta} = k$), one from a student from group A and one from a student from group B, the expected aptitude of a group A applicant will be higher than that of a group B applicant. This can be true even though the noisy signal $\hat{\theta}$ is an "unbiased" estimate of θ for both groups in that its average value equals the true unobserved value for both groups of applicants. The reason this can occur is simply that the aptitude prediction is based on the noisy signal, $\hat{\theta}$, as well as the base rate, $\overline{\theta}$, and we have assumed that group A members are on average higher performing.

The conclusion that some draw from these observations is that discrimination is simply an unfortunate by-product of rational information processing when signals are noisy. Given two equally promising-looking law school applicants, one from group A and one from group B, the admissions officer's prediction will be that the group A applicant's aptitude is higher. This might appear to be an unfounded prejudice against the group B applicant – we have stipulated that their file is exactly as good as that of the group A applicant, so why is the decision-maker looking more favorably upon the latter? But the point of the Phelps/Arrow approach is to show that these prejudices may not in fact be unfounded. They are founded, on this model, on the decision-maker's prior observations of patterns in the track record of members of groups A and B. Given that the data from the application file are noisy, one might argue that it is reasonable for a decision-maker to take these track records into account and use them to contextualize the noisy applications when when forming estimates of applicants' aptitudes.

The main problem with the statistical theory of discrimination is that it attempts to absolve the decision-maker of moral turpitude by assuming that the two groups *in fact* differ in their true aptitude. In other words, it assumes that group B members are genuinely (on average) inferior with respect to θ. Aigner and Cain (1977) highlight the strangeness of this assumption in a theory of discrimination: it seems to stack the deck against any conclusion that discrimination is wrong. The approach instead shows us that someone who uses their knowledge of group B's inferiority would on average predict that group B members will fare less well. And indeed they would – because their prediction is a weighted average of the group mean (i.e., the base rate) and the noisy signal: $p\hat{\theta}_i + (1-p)\overline{\theta}_i$. But what remains unaddressed by the statistical theory of discrimination is precisely what matters most to questions of bias and fairness: one would think that bias is a matter of whether people who *resemble* one another in all relevant respects are treated differently, and not of whether people who genuinely *differ* in some relevant respects are accordingly treated differently.

Moreover, from a Strawsonian point of view, this approach is *looking at the wrong attitudes*. It attempts to locate bias in the decision-maker's beliefs, but that is not (or, at least, not primarily) where one would expect to find it. The approach tells us nothing about whether a decision-maker is prejudiced, or bigoted, or is the bearer of racial animosity or resentment – in other words, about the degrees of good or ill will that she harbors toward members of groups A and B. Having observed that the decision-maker's estimates of aptitude vary systematically by group, one wants to know: How do they feel about this? Are they pleased to have an excuse to admit members of group B at a lower rate? Or are they reluctant, and sympathetic to the plight of the equally-good-on-paper group B applicants? Or are they entirely indifferent? The answers to these questions tell us how the decision-maker *regards* members of different groups, allowing us to glean insight into their non-doxastic attitudes and thereby to assess their degree of bias. And these questions are accordingly similar to the ones that typically arise in legal disputes over the constitutionality of allegedly discriminatory policies in the context of the Fourteenth Amendment's Equal Protection Clause in the United States (under the banner of discriminatory purpose),[17] as opposed to the questions that arise in the narrower statutory disputes in which disparate impact is directly legally prohibited.[18] More generally, what is of interest from a moral point of view is our ability to infer something about whether the decision-maker holds groups A and B in equally high regard and has equal concern for their interests.

One might further wonder how a decision-maker feels about the fact that, on

[17] *Washington v. Davis*, 426 U.S. 229 (1976). See also, *Personnel Administrator of Massachusetts v Feeney*, 442 U.S. 256 (1979) (noting that laws violating the Foureenth Amendment's Equal Protection Clause are passed because of, not merely in spite of, their adverse effects upon an identifiable group).

[18] For example, *Griggs v Duke Power Co.*, 401 U.S. 424 (1971), interprets Title VII of the Civil Rights Act of 1964 to prohibit employment practices with a racially disparate impact.

the Phelps/Arrow approach, they are making predictions about the aptitude of particular *individuals* based on the average aptitude of a social group to which these individuals belong. Indeed, the Phelps/Arrow approach implies that the probability of an A group candidate being admitted to our law school is overall greater than the probability of a B group candidate being admitted:

$$\Pr(\text{Admit}|A) > \Pr(\text{Admit}|B)$$

because $\overline{\theta}_A > \overline{\theta}_B$. We might worry that this seems unfair to B group applicants; every individual member is handicapped vis-a-vis law school admissions by virtue of their membership in the group, which looks like a particularly crude form of stereotyping.

This raises a thorny problem, which is often referred to as the problem of "naked statistical evidence" in law and philosophy.[19] The problem in a nutshell is that, according to the statistical theory of discrimination, predictions about individuals ought *rationally* to be based in part on base rates of the relevant trait(s) among groups to which the individuals belong. But philosophers and legal scholars do not automatically assume that what is rational is also moral, and a growing body of literature also challenges the rationality of this approach. Thus we can ask two questions: (1) Does rational information processing really require such adverse inferences against minority applicants? And (2) Even if it does, is that fair? If the answers are "yes" and "no" respectively then one might think – as Gendler (2011); Basu (2019); Basu and Schroeder (2019) have argued – that there is an unavoidable conflict between the requirements of rationality and those of morality. However, Johnson King and Babic (2020) have argued that the morality and rationality of the sorts of pernicious predictive inferences licensed by the Phelps/Arrow approach cannot be assessed separately from one another, as there is no such thing as a value-free inferential process (see also Johnson, 2023, for a similar argument). On their approach, which we further develop here, a decision-maker's credal states *themselves* manifest her evaluative attitudes. This means that the decision-maker's credences can be morally assessed in the course of assessing their rationality. And it means – *pace* economic theories of efficient discrimination – that, even if groups A and B do in fact differ in their true aptitude, predicting accordingly does not necessarily absolve the decision-maker of moral turpitude. We will spell this out in more detail below.

3.2 Discriminating values

Unlike Phelps and Arrow, Becker (1957) develops a notion of discrimination in employment that locates discriminatory attitudes in a decision-maker's utility

[19] This problem was first addressed in the late 1960s and early 1970s, including by Kaplan (1968), and Tribe (1971). Then in the 1980s, including by Cohen (1981), Nesson (1985), and Thomson (1986). And more recently, with Colyvan et al. (2001), Schauer (2003), Redmayne (2008), Buchak (2014), and Cheng (2013). Most recently, philosophers such as Basu (2019) and Moss (2018) have written on this issue as well.

function – hence the name "taste-based discrimination". On Becker's approach, a non-discriminating employer seeks simply to maximize her profit: her utility function is a product of how many outputs a worker can produce multiplied by the price she can charge for each output, minus the amount she has to pay the worker for their time. Meanwhile, a discriminating employer has an additional argument in their utility function, which Becker termed the *coefficient of discrimination*. This measures the extent of the decision-maker's aversion to hiring employees from a certain group; we may think of it as a measure of her degree of animosity toward the group. Call it d.

Prejudiced employers, for whom $d > 0$, therefore behave as if the salary they have to pay minority applicants is $c + d$ instead of just c, where c represents the per-worker cost. Effectively, the decision-maker's ill will toward the minority group leads her to regard hiring an individual from this group as more "costly" than hiring a member of the majority group at the same wage. Quite apart from the financial costs, then, hiring minorities is something to which she has an aversion, quantified by the magnitude of her coefficient of discrimination.

Economists have criticized Becker's theory because it adds an argument to the utility function – tinkering with the math in this way can allow one to reach any desired conclusion. But, from our perspective, Becker's approach offers two fundamental insights. First, discriminatory attitudes are non-doxastic: they must ultimately be traced to the decision-maker's values (i.e., what she cares about). Within a Bayesian approach, this means that assessments of bias and fairness must examine the decision-maker's utility function and the discriminatory attitudes to be found therein.

The second insight is that bias and fairness are graded concepts. In Becker's approach d can take any real value. If $d > 0$ with respect to group B, then the decision-maker is biased against group B. If $d < 0$, then the decision-maker is biased in favor of group B – i.e., she derives some kind of enjoyment from hiring group B workers, or sees doing so as valuable, to a degree that can be quantified in terms of a positive dollar equivalent. Therefore, if we have two decision makers k and l, and $d_k > d_l$ with respect to group B, then k is more strongly biased against group B: it would cost more to incentivize k to hire a group B worker than it would to incentivize l to do the same. Taking these lessons from Becker, our approach will be both utility-based and graded.

Now, recall the lesson that we drew earlier from the phenomenon of moral luck: whether a decision-maker holds different individuals or groups in equal regard is not a matter of the outcomes that she happens to bring about, but of what she cares about. This lesson can easily be implemented within a utility-based approach to evaluating discrimination. That is because a rational decision-maker's utilities are precisely where her values are represented.[20]

[20] Recall also Smith's point that evaluative judgments need not be propositional attitudes and may simply be "tendencies to regard certain things as having evaluative significance";

It is an agent's utility function that determines whether she cares about any two things to the same degree or cares about one more than the other. And it is thus the utility function that determines whether she cares about the interests of different individuals and/or groups to the same degree or cares about one more than the other. This point also makes it easy to see how fairness and bias will be graded phenomena on a utility-based approach; if one manifests equal regard for two individuals by caring about their interests to an exactly equal degree, then one likewise displays bias in favor of one individual and against another *to the extent that* she cares about the interests of the former more than those of the latter.

Indeed, the utility function is the *only* place where one might locate a decision-maker's degrees of concern or regard in a manner that can apply not only to human agents but also to algorithms – including black-box algorithms. In Becker's employment example, the utility function is a representation; it models the discriminating employer's *aversion* to hiring employees from one group, which is a contentful mental state that the employer is in. This dovetails nicely with a popular recent development of the quality of will approach that locates agents' good and ill will in our intrinsic desires and aversions (Arpaly and Schroeder, 2013). On this approach, desire and aversion are mental states realized in the human brain's reward and punishment systems, with a wide range of effects – they influence our patterns of attention, give rise to motivations to act, and lead us to feel pleasure and pain when our estimation of the chance of an outcome that we desire or to which we are averse increases or decreases. But none of *that*could possibly apply to algorithms. Algorithms' quality of will cannot be realized in the brain's reward and punishment systems because algorithms do not have brains. Indeed, as we observed earlier, algorithms do not even have contentful mental states in the usual sense. Nor do they display patterns of attention, experience motivation, or feel pleasure and pain. If we are to ascribe good and ill will to a decision-making algorithm in a way that is not unduly anthropomorphic, then, we cannot understand the utility function as a mere representation. Instead, the utility function is precisely what the algorithm's degrees of concern or regard *consist in*; there is no more (and no less) to its will – be it good or ill – than this. This means that examining an algorithm's degree of bias requires us to estimate the utility function that its decisions are optimizing.

To understand our view, return briefly to the Phelps/Arrow approach. We saw earlier that if base rates really do differ between groups A and B, then the average prediction for an individual from group A will differ from the average prediction for an individual from group B. But what does it mean for base rates to *really* differ? As Babic et al. (2021) explain, it is entirely possible that our law school admissions officers will come to *falsely* believe that base rates of aptitude differ across groups A and B simply because group B members have historically

whatever exactly utilities might be, it seems clear that they are, at a minimum, ways of regarding certain things as having evaluative significance.

been discriminated against in the legal profession, facing many barriers to their professional success – including implicit biases on the part of their superiors as well as explicit obstacles – which collectively lead to actual discrepancies in performance between group A and B that then convey the misleading impression that group A members have on average higher aptitude. While this sort of thing is not logically guaranteed to occur, it does occur in just about every historical example of discrimination. When a racial, gender, ethnic, etc. group faces systemic prejudice and institutional barriers, members of the group ultimately perform less well on some standard measures of success.

This means that, when observing trends in the performance of admitted students over the years, the law school admissions officer is not really answering the question: What is this student's true aptitude, call it θ_{real}? Rather, they are answering the question: What is this student's aptitude in a world that systematically favors people from group A, call it θ_{pseudo}? The former question would be exceedingly hard to answer, since we can only observe lawyers in the particular society in which we find ourselves.[21] But surely the latter is not an appropriate basis on which to make admissions decisions. When we first introduced the notion of aptitude, its attractiveness as a criterion for admission to law school depended on our interpreting it as some kind of "true" aptitude.

Thus, while it is correct that the law school admissions officer may have observed different base rates about θ, this is only so of a θ that should not be relevant to law school admissions decisions. We do not have good information about the θ that we really care about (θ_{real}). Once we realize this, we might then start with a uniform prior for θ and make our decisions entirely on the basis of $\hat{\theta}$ – i.e., the application file. That is, the difference between a group A member i and a group B member j can be given entirely by the difference in the quality of their applications:

$$(p\overline{\theta}_A + (1-p)\hat{\theta}_i) - (p\overline{\theta}_B + (1-p)\hat{\theta}_j) = (1-p)(\hat{\theta}_i - \hat{\theta}_j) \propto \hat{\theta}_i - \hat{\theta}_j.$$

This would avoid the kind of discrimination predicted by the Phelps/Arrow model, which is also the kind that motivates worries about naked statistical evidence in law and philosophy – i.e., about base rates skewing estimates against disadvantaged applicants. However, it does not solve all our problems. Even if we ignore groups' historical base rates and make admissions decisions solely on the strength of each candidate's file, it can easily happen that disproportionately few members of the B group – the group that has been discriminated against in the legal profession – are admitted. For this group will likely also face various obstacles in society at large, such that a child growing up in group B endures socioeconomic hardships and confronts educational barriers that, on average,

[21] To make inferences about θ_{real} on the basis of θ_{pseudo} we would have to estimate the proportion of high aptitude minority group lawyers who fail to succeed, and the proportion of low aptitude minority group lawyers who succeed. Babic et al. (2021) propose a method for making this estimate for binary data.

lead to differences between group A and group B application files. That is, $\mathrm{E}[\hat{\theta}_A] > \mathrm{E}[\hat{\theta}_B]$.[22]

If the admissions officer believes that one group's application files are indeed less informative than the other, then she may want to give members of that group the benefit of the doubt in her admissions decisions. In other words, a kind of *evidential affirmative action* might be warranted. While this might look like a kind of unfairness – a procedural inequality as against the majority group – our approach allows for the possibility that it is morally required.

A natural way to give the benefit of the doubt to a disadvantaged group is to structure one's prior in a way that privileges members of that group on the basis of moral considerations. This is precisely what Johnson King and Babic (2020) argue for. In particular, an agent should choose a prior for θ in a way that appropriately reflects her attitudes to the relative costs of false positive and false negative mistakes in her decision problem. While these costs could be taken to be equal, that would not be a way of avoiding making judgments about their relative magnitude but would instead be a particular such judgment. In the case we are considering, though, the admissions officer might well *not* take the costs to be equal. She might be more worried about failing to admit high-aptitude group B applicants (a false negative mistake) than admitting low-aptitude group B applicants (a false positive mistake). This could be either to rectify the injustice done to the group by their having been historically discriminated against, or to do justice to particular group members of extraordinary talent – since they will be particularly difficult to identify and yet particularly important to identify.

Here, then, is how we can use these attitudes toward epistemic risk (Babic, 2019) to identify a reasonable prior. First, let us suppose that what we observe are lawyers who succeed, $X = 1$, and lawyers who do not, $X = 0$. This is a little fanciful, but not too much: for example, sticking with the earlier heuristic, we might say that those who pass the bar "succeed". We know that higher-aptitude lawyers are more likely to succeed than their lower-aptitude peers. So let θ follow a beta distribution with parameters α and β, so that $\mathrm{E}[\theta] = \alpha/(\alpha + \beta)$. And let $s(p, I_{X_i})$ be the decision-maker's scoring rule for the probability assigned to the proposition/event X_i, $p = Pr(X_i)$ – for example, representing the event that a particular applicant would succeed – where $I_X = 1$ if X is true and 0 otherwise.[23] A *symmetric* scoring rule is indifferent between approaching

[22] One might think that if $\mathrm{E}[\hat{\theta}_A] > \mathrm{E}[\hat{\theta}_B]$ then it must also be true that $\overline{\theta}_A > \overline{\theta}_B$, since $\mathrm{E}[\hat{\theta}] = \theta$. But that is not necessarily so, because $\mathrm{E}[\hat{\theta}] = \theta$ only if the variances between θ_A and θ_B are the same. In the example we are currently considering, variances may not be identical. For instance, we can quite easily imagine a situation where the minority group B is on average disadvantaged – due to, say, socioeconomic and educational barriers – while containing many extraordinary unrecognized superstars, while group A, exploiting all of their advantage in a fairly uniform manner, is densely concentrated around a point of mediocrity.

[23] $s(p, I_{X_i})$ should be continuous, and $s(p, 1)$ should be monotonically decreasing while $s(p, 0)$ should be monotonically increasing.

inaccuracy in the false positive direction and in the false negative direction – i.e., $s(p,1) = s(1-p,0)$ for all p – whereas an *asymmetric* scoring rule is not. Johnson King and Babic (2020) require that $\mathrm{E}[\theta] = p^*$ where p^* satisfies $s(p,1) = s(p,0)$, the point of minimum epistemic risk.[24] But, since $\mathrm{E}[\theta] = \alpha/(\alpha+\beta)$, this framework imposes a requirement on the permissible values of α and β. These values are in turn determined by the decision-maker's attitudes toward epistemic risk with respect to X.

If we are giving a student from group B the benefit of the doubt, then for that group $p^* > 0.5$, and so $\alpha/(\alpha+\beta)$ must be greater than 0.5. When we observe the application file, $\hat{\theta}$, our posterior will then be a weighted average of this prior and the evidence – i.e., $\frac{c\alpha}{(\alpha+\beta)} + (1-c)\hat{\theta}$ – where the weight c given to the prior increases in $(\alpha+\beta)$. This is the same expression we saw earlier in the Phelps/Arrow model (pg. 13), except that $\overline{\theta}$ is no longer a naive base rate; it is rather a prior that is structured by taking moral considerations into account. In other words, the prior is constructed by taking account of the costs of error with respect to admitting a low-aptitude applicant and failing to admit a high-aptitude applicant, given that the applicant is a member of disadvantaged group B. The higher the sum of α and β is, the less weight the decision-maker will give to the application file. And the higher $\alpha/(\alpha+\beta)$ is, the more benefit of the doubt the decision-maker is giving to the applicant.

This approach to fairness is ultimately in the spirit of Becker's taste-based approach, since p^* determines the appropriate prior and p^* depends on the symmetry or (degree of) asymmetry of the agent's scoring rule – which, in turn, depends on how averse she is to false positive and false negative mistakes. It is also very much in the spirit of Strawson's approach. Strawson wrote that "[w]e should consider... in how much of our behaviour the benefit or injury resides mainly or entirely in the manifestation of attitude itself. So it is with good manners, and much of what we call kindness, on the one hand; with deliberate rudeness, studied indifference, or insult on the other" (Strawson, 1982, p.63). As we have noted, a symmetric scoring rule is one that is *indifferent* between approaching error in the false positive direction and in the false negative direction. As a substantive moral matter, this indifference may be an inappropriate response to the actual costs of false positive and false negative mistakes; in our law school admissions example, for instance, the admissions officer may have substantive moral reason to worry about the files of group B applicants being particularly noisy due to prior socioeconomic and prejudicial harms' negatively impacting group members' ability to develop a file that reflects their true aptitude. Under such circumstances we might assume that it is inappropriate to be indifferent between false positive and false negative mistakes: the latter error is worse, since it perpetuates group-based injustice. In that case, in Strawsonian

[24] In this project, we articulate p^* in terms of where $s(p,1)$ intersects with $s(p,0)$ – i.e., the point where there is no accuracy uncertainty, hence the point of zero epistemic risk. More generally, p^* is the minimum of the formal epistemic risk function, as articulated in Babic (2019).

spirit, we may react to an admissions officer's "studied indifference" between the two types of error with indignation at their lack of concern for the interests of group B and its members. Equally and oppositely, we might see an admissions officer's adoption of an asymmetric scoring rule – one that manifests a greater degree of aversion to false negative mistakes – as a form of "what we call kindness". This is an algorithmic version of precisely the sort of case that Strawson was talking about, wherein the benefit or injury done to an individual resides primarily in a decision-maker's attitudes – though, *pace* Strawson, it does not reside "entirely" there, since these are attitudes that can have significant material import for the lives of those whose fates are thereby decided.

Now we are in a position to spell out exactly how our approach applies to black-box algorithms. First, let us draw out an implication of the preceding summary of Johnson King and Babic (2020): on their approach, there is a clear sense in which mistakes (like fairness itself) are *graded*. The law school admissions officer's estimate of each applicant's aptitude will take a real value, as will the applicant's actual aptitude. The bigger the discrepancy between the estimate of a student's aptitude and the student's true aptitude, the greater the mistake. If the difference is positive, so that the estimate of aptitude is higher than the student's true aptitude, then the mistake is in the false positive direction. And if the difference is negative, so that the estimate is lower than the student's true aptitude, then the mistake is in the false negative direction. Hence we are able to speak of *approaching* mistakes in the false positive and false negative direction, rather than simply of making false positive and false negative mistakes. The attitudes to error used to structure a decision-maker's prior can thus be defined in terms of a continuous epistemic utility function, since, for every small change in credal value, the agent can assess the cost of moving away from the truth against the cost of moving toward it. However, they may not be indifferent between the two directions of mistake; for example, as we have noted, they may reasonably worry more about failing to admit high-aptitude students than falsely admitting low-aptitude students. These considerations pave the way for a more direct way to incorporate values into the decision problem, which is generally equivalent to that developed in Johnson King and Babic (2020) but more useful for assessing decision-making algorithms.

Let us suppose that our utility is given by the loss function, $L(\theta, \hat{\theta})$, such that we suffer cost $(\hat{\theta} - \theta)k_0$ when $\theta < \hat{\theta}$ and $(\hat{\theta} - \theta)k_1$ when $\theta > \hat{\theta}$. As Bayesians, we want to minimize the posterior expected loss, $\mathrm{E}[L(\theta, \hat{\theta})]$. The derivative of this expression is given by:

$$\frac{\partial}{\partial \hat{\theta}} \mathrm{E}[L(\theta, \hat{\theta})] = k_0 F(\hat{\theta}) - k_1[1 - F(\hat{\theta})],$$

where F is the cumulative distribution function. Setting this quantity to 0, we obtain:

$$F(\hat{\theta}) = \frac{k_1}{k_0 + k_1}.$$

And finally,

$$\hat{\theta} = F^{-1}\left(\frac{k_0}{k_0 + k_1}\right),$$

where F^{-1} is the fractile function. What this means is that, when we are differently worried about approaching error in the false positive direction than in the false negative direction, our best guess for an applicant's aptitude should be a fractile determined by our relative costs of error.

Notice that if $k_0 = k_1$ then our best guess ought to be the median of the posterior distribution for θ. But using the median as our best guess is not a mathematical inevitability. It is instead the solution to an optimization problem for a decision-maker who is precisely indifferent between approaching false positive and false negative mistakes. But this attitude to error is something that the decision-maker brings into the decision problem and not something that they are forced to adopt by the structure of the problem itself. As such, the decision-maker's attitudes – what sorts of losses they care about, and to what degree – are morally assessable. If we find that the decision-maker uses one fractile for, say, black applicants and another fractile for white applicants – i.e., k varies by race – this means that they value false positive and false negative mistakes differently for applicants of different races, since the degree to which they are more averse to one type of error than another varies by racial group. And it is this discrepancy in the degree to which the decision-maker cares about different groups' interests, we say, that constitutes *prima facie* unfairness.

Why *prima facie*? Because it is not automatic that a decision-maker with different decision thresholds for different groups *ipso facto* cares about the groups' interests to different degrees. On the contrary, there can be cases in which showing equal regard for each individual *requires* having different thresholds for different groups – as in the case where the law school admissions officer has reason to expect group B's signals to be particularly noisy in light of the pervasive socioeconomic and interpersonal disadvantages they face, and may thus adopt a lower decision threshold for group B applicants both with an eye toward rectifying the surrounding injustices perpetrated against the group and with an eye to doing right by the high-aptitude group B individuals who face partiuclar difficulties in conveying their true aptitude. Hence differential decision thresholds can turn out to be *ultima facie* supported by fairness considerations. This means that to identify such differences in threshold is not necessarily to indict a decision-maker of bias. Instead, it is to observe that the decision-maker cares differently about errors across certain groups and thereby to open up a discussion about whether this discrepancy can be given a substantive moral rationale.[25] Other things equal, caring differently about errors across groups

[25] Conversational models of responsibility are popular in the post-Strawson literature – see, for instance, McKenna (2012). The basic idea is that a blamer calls upon the blamee to either justify their action (in this case by showing that it is, contrary to appearances, supported by fairness considerations) or apologize and make amends.

amounts to bias in favor of one group and against the other(s). But the same discrepancies can be morally appropriate when other things are not equal.

We can envision all sorts of ways that this sort of substantive discussion, engendered by an identification of a differential decision threshold, might go. For instance, in the foregoing we assumed that an admissions officer may reasonably be more concerned with failing to admit high-aptitude students than with admitting low-aptitude students, but the reverse could also be true. The admissions officer may realize that their institution does not have adequate support structures in place to enable students who are struggling to get back on track. The institution may accordingly have high rates of attrition. In that case, the admissions officer may reasonably worry about falsely admitting low-aptitude students on the grounds that their admission might be seriously detrimental to them, forcing them to incur significant financial costs for a series of classes with which they cannot keep up and that leads to major stress and feelings of inadequacy. Thus the officer might reasonably think that their lower-aptitude applicants would be better-off attending a different institution where they can thrive. And so they may be equally or more concerned about falsely admitting low-aptitude students than failing to admit high-aptitude students, again on reasonable grounds.

Alternatively, when it comes to the historically-discriminated-against group B, the admissions officer may think that a policy of "studied indifference" *is* the way to show concern for this group members' interests. For they may worry that a differential decision threshold would create the impression that the admitted students from group B were admitted primarily for the sake of enhancing campus diversity and not on the basis of their knowledge and skills. The admissions officer may reasonably worry that this, too, would lead to feelings of inadequacy among group B students – not to mention the potential for patronizing differential treatment of these students on the part of their instructors. And they may think that explicitly "holding everybody to the same standard" is the best way to mitigate against this demeaning impression and its deleterious effects. In short, the moral costs associated with false positives and false negatives in interesting and important social decisions are rarely clear-cut. Our point is not to take a stand on the actual relative magnitude of the costs involved in college admissions or any other important social decision. The point is rather that we can identify differential decision thresholds as evidence of *prima facie* unfairness and in doing so can open up a normative discussion about what the costs are and what the thresholds ought to be.

Our basic idea can be simplified further by focusing on binary choice problems with at least partially observable binary outcomes – the kind of case that is typically presupposed in the literature. For example, loans are either approved or denied, and, at least of those which are approved, we observe who pays back and who does not.[26] Parole is either granted or not, and, at least in popular

[26]The question of how to approach what would happen to those who are denied a loan is

data sets such as COMPAS, we observe all the outcomes. And so forth. In these contexts it makes sense to speak not only of approaching error in the false positive and false negative directions but also of categorically making false positive and false negative mistakes.

In such cases, the loss function is simpler still. Suppose we have to choose between extending and not extending a loan, under uncertainty over whether the recipient will repay with probability p. We might let b_1 represent the benefit if the loan is extended and repaid and let b_2 represent the benefit if the loan is not extended to someone who would otherwise default. Similarly, let $-c_1$ represent the cost of extending the loan to someone who would default (a false positive mistake, and $-c_2$ the cost of refusing the loan to someone who would repay (a false negative mistake). Plausibly, $b_1 > b_2$ and $-c_1 < -c_2$. The loss function then looks like this:

	p	$1-p$
	Repay	Default
Extend	b_1	$-c_1$
Do not extend	$-c_2$	b_2

The expected value of extending the loan is $pb_1 - (1-p)c_1$. And the expected value of denying the loan is $(1-p)b_2 - pc_2$. Hence we should extend the loan if

$$\begin{aligned} & pb_1 - (1-p)c_1 > (1-p)b_2 - pc_2 \\ \rightarrow & pb_1 - c_1 + pc_1 > b_2 - pb_2 - pc_2 \\ \rightarrow & pb_1 + pc_1 + pb_2 + pc_2 > c_1 + b_2 \\ \rightarrow & p > \frac{c_1 + b_2}{c_1 + b_2 + b_1 + c_2} \end{aligned}$$

To simplify further, we might let $b_1 = b_2 = 0$, in which case we should extend the loan if $p > \frac{c_1}{c_1+c_2}$. Letting $t = \frac{c_1}{c_2+c_2}$, this means that we should extend the loan if the probability of repayment exceeds a threshold t, which is determined by the extent to which we care about false positive and false negative mistakes (and in the more general case also true positive and true negative decisions). In other words, the threshold, t, is directly and fully determined by the decision maker's utility function: $t = f(c_1, c_2, b_1, b_2)$. Therefore, if the decision threshold varies by racial, ethnic, gender, or other groups, then the decision-maker has a different utility function depending on whether the relevant individual is black vs. white, male vs. female, citizen vs. alien, or whatever the case may be. In other words, a difference in decision thresholds is squarely a case of taste-based discrimination in the form of Becker.

Now, under the approach developed in Johnson King and Babic (2020), the decision-makers' values are located in their priors, rather than in their loss

an interesting censored data problem, but we set it aside here.

or utility functions (as evidenced by decision thresholds). But, as has been observed, the prior and loss are dual to each other.[27] In Johnson King and Babic (2020), the predictive probability is given by $\alpha/(\alpha+\beta)$, which are themselves in part determined by the agent's attitudes to error. So now suppose that $\alpha/(\alpha+\beta) = t$ (i.e., one particular individual is, say, denied admission because they are exactly at the threshold, and admission requires that $p > t$). There are two ways to explain this. We can say the following: the individual is just below the threshold, the prior is non-informative (uniform) (i.e. $\alpha = \beta = 1$), but the threshold incorporates the agent's attitudes to error (i.e. $c_1 \neq c_2$). Or, we can say the following: the individual is just below the threshold, the prior incorporates the agent's attitudes to error (i.e., $\alpha \neq \beta, \alpha > 1, \beta > 1$) but the threshold is value-free (i.e, $c_1 = c_2 = 1$). These are equivalent representations of what has happened to a particular applicant.

The point here is that, due to this duality between the prior and the decision threshold, we can explain a decision by saying either that the prior is value-laden and the threshold is neutral or that the prior is neutral and the threshold is value-laden. And, of course, we can also say that values affect both the loss and the prior. It is also possible to maintain that values *should* affect neither the loss nor the prior. As Johnson King and Babic (2020) argue, however, the latter is not a morally "neutral" position but rather a particular attitude toward the moral costs of error.

Our view in this project is not to argue for any specific attitude, but simply to state the following:

> *If* we find that a decision maker's threshold differs by group, *then* we have *prima facie* unfair treatment of one group, and the extent of that unfairness can be quantified by the extent of the difference in thresholds, which in turn corresponds to a difference in the costs and benefits associated with the correct/incorrect decisions for members' of different groups.

This is unfairness in Becker's sense, but it is also a way of thinking about bias and fairness that can apply to algorithms and can be given a robust philosophical rationale, drawing on the theoretical resources of the quality of will approach to explain why we care so much about the attitudes toward the costs of error embodied by a decision-making algorithm. To repeat: we care about decision-makers' utilities because we care about what decision-makers care about – we care whether they care about the right things, and to the right degrees. We do not care about observed disparities in outcomes for their own sake. Those can occur for many reasons, only some of which are connected to underlying discriminatory attitudes.

[27] Robert (2007) use this expression, but what is meant should be clear from the relationship between α, β and t in our model, as we explain.

One important upshot of our approach is that it is not possible to shy away from challenging ethical and political theorizing when diagnosing algorithmic bias. For *prima facie* unfairness may not be *ultima facie* unfairness, and the upshot of a diagnosis of the former is that we must engage in ethical and political reasoning to determine whether or not we have a case of the latter. Here it is important to remember that decision-making algorithms are designed, owned, and used by humans. When we observe *prima facie* unfairness in an algorithm's outputs, it is the humans in charge of it who can be called upon either to justify these discrepancies with substantive moral argument or – if the discrepancies are unjustified – to change something about the algorithm's design or their use of it in decision-making. This is why we say that identifying *prima facie* unfairness opens up a discussion: we can approach the humans in charge of the algorithm, inform them of the discrepancy in its degrees of concern for the interests of different groups, and ask what they intend to do about it. For someone who genuinely does harbor equal concern for the different groups, this information about the problem and the ensuing conversation in search of a solution would presumably be welcome.

4 Revisiting thresholds

We mentioned above that this simple idea – namely, that a rational Bayesian maximizing expected utility will make binary choices on the basis of a probability threshold – appears in several recent articles on fairness and discrimination. In this section, we relate those articles to our project in more detail.

As explained earlier, while Simoiu et al. (2017) similarly use a threshold (in their case, to evaluate police racial discrimination in vehicle stops), they quite consciously do not attempt to offer substantive moral or political arguments to defend the assumption that a difference in decision thresholds is evidence of unfairness or discrimination. Rather, they explicitly state that this is an assumption their project makes. They then go on to develop a statistical model for estimating thresholds from observations of police vehicle stops, search rates, and hit rates (i.e., proportions of searches where illegal items are found). That is, their model takes as its inputs the available information on individual police stops and produces as its output a search threshold by the driver's race and the stopping officer's department.

In Kleinberg et al. (2018) the framing is as follows. There is, on the one hand, an *efficient planner*, akin to our law school admissions officer whose goal is to admit the "best" students. In our notation, they choose students so as to maximize $\mathrm{E}[\theta_A + \theta_B]$ among the admitted class. So, if $\overline{\theta}_A >> \overline{\theta}_B$, then the optimal class for an efficient planner would have many more (perhaps all) group A students. By comparison, the *equitable planner* seeks to admit the group of students with the highest estimated aptitude subject to a further condition: that the class selected is sufficiently diverse. For example, they may seek to maximize $\mathrm{E}[\theta_A + \theta_B]$ subject to the constraint that $\#A/\#B < 1$ (i.e., at least

50 percent of students are from the minority group B).

Kleinberg et al. (2018) show that something analogous to a threshold decision rule is optimal for the efficient planner: the efficient planner should simply rank students by predicted aptitude and select the n best students – the threshold t would in this case correspond to the aptitude of the marginal student. The authors then show that the equitable planner should do something similar: supposing for illustration that they must admit 100 students, but they seek to admit 70 from group A and 30 from group B, then the equitable planner should simply rank all group A students, admitting the best 70, and rank all group B students, admitting the best 30.[28] This implies that the threshold for admitting a group A member can be higher than the threshold for admitting a group B member when, for example, $\overline{\theta}_A >> \overline{\theta}_B$. Notice further that, in order to implement such a procedure, one would need to explicitly use group identity in the admissions process. Hence the equitable planner implements a kind of affirmative action policy.

However, Kleinberg *et al* do not explain *why* fairness might require a lower threshold for applicants from certain groups as compared to others. It is presupposed *ex ante* that the equitable planner is a fairer planner than the efficient planner and that the equitable planner cares about diversity – to a particular precise extent, given that they are able to specify a precise quota of group B applicants that they seek to admit. But no explanation is given for the assumption that this is what fairness requires. And the assumption can be challenged: it is far from clear that fairness requires simple quotas. Indeed, from a legal perspective, the Supreme Court has made clear that simple racial quotas are in fact unconstitutional[29] and the pending Harvard University admissions litigation[30] could go further and find that even soft quotas are likewise unconstitutional.

In short, while these authors explore the trade-off between fairness and efficiency, what is left unsaid, as in Simoiu et al. (2017), is why fairness requires behaving as the "equitable" planner does (i.e., using different thresholds for different groups and implementing an affirmative action policy – a very controversial position). This is precisely the question that our project takes up.

[28] Similarly, Corbett-Davies et al. (2017) argue that if one takes, as a persuasive conception of fairness, certain of the definitions that exist in the current literature – such as predictive disparity – then one can likewise express fairness constrained optimal decision making in terms of a threshold. This is not surprising given our discussion so far. Adding a constraint is equivalent to changing the utility function. The rational Bayesian continues to decide based on a threshold, but the particular value of that threshold changes. However, these authors do not say anything about which particular fairness constraint is normatively defensible, or why. This is our project

[29] See, for instance, Justice Powell's infamous argument against quotas in *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978), and more recently *Grutter v. Bollinger*, 539 U.S. 306 (2003) (affirming the unconstitutionality of racial quotas).

[30] *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College*, Docket 20–1199.

We have explained how the threshold embodies the decision-maker's degrees of concern or regard for different groups, which, from a Strawsonian perspective, is precisely what we are interested in when we are interested in fairness. When these attitudes differ across groups, the decision-maker manifests unequal regard – something to which we appropriately respond with resentment and indignation in and of itself, quite apart from the further material harm to which unequal regard can lead when it is manifested by those who make decisions with substantial material implications for others' lives. We have tied together Becker and Strawson via Bayes.

On our approach, the answer to what fairness requires cannot be read off an equation. This is because morality cannot arise out of mathematics alone. Once *prima facie* unfairness has been identified, it behooves the decision-maker (or the humans in charge of it) to reflect on the appropriate costs of error in their decision problem and decide whether showing equal concern for everyone's interests in this case requires equal thresholds or not. And, of course, others may disagree with the decision-maker on precisely this point, which is when the real conversation begins. This is, in every case, a substantive moral question that can only be answered with substantive moral argument.

Notice that the most obvious question about our approach – But how do you decide which threshold is right? – mirrors the most obvious question about Bayesianism in general – But how do you decide which prior is correct? – precisely because of the duality between the loss and the prior. And, ultimately, it is misguided to expect a mathematical answer to such a question. The Bayesian approach invites the researcher to justify her selection of a prior. Similarly, our approach to fairness invites the decision-maker to justify the costs and benefits that inform her threshold.

At this point someone might worry: is our approach just another version of the outcome-based approaches that we criticized at the outset? After all, if we are going to estimate thresholds, how else can we do it other than through analysis of observed outcomes? We earlier argued that measures of fairness focused on outcomes all fall prey to the problem of moral luck, failing to acknowledge the fact that the outcomes an agent happens to bring about are only very loose indicators of what she cares about. Moreover, we argued that whether a decision-maker has equal regard for the interests of different individuals or groups is a matter of what she cares about rather than of what she brings about. But we are now partially endorsing some models that also base their judgments on an analysis of the outcomes that the decision-maker brings about (as Simoiu et al. (2017) do). So isn't our approach also outcome-based? And, if so, doesn't it fall at the same hurdle?

The answer is that there are two things that one might mean by "based" when one says that an approach is "outcome-based", one of which is problematic but not true of our approach whereas the other is true of our approach

but unproblematic. A way of assessing the fairness of decision-making algorithms could be based on the algorithm's outcomes in either a *metaphysical* or an *epistemological* sense. It is outcome-based in a metaphysical sense if it takes disparate outcomes to constitute unfairness; that is, if it holds that what *makes* an algorithm unfair is that the algorithm produces disparate outcomes for individuals or groups. This falls prey to the moral luck objection.

But a way of assessing the fairness of decision-making algorithms could also be based on outcomes in a merely epistemological sense, if it uses the algorithm's outcomes as evidence of whatever it is that actually determines fairness. Indeed, as we explained above, this is typically how courts evaluate discrimination in the context of Equal Protection jurisprudence, treating outcomes as relevant insofar as they speak to discriminatory intent or purpose. And this is also true of our approach. We hold that outcomes are not determinants of fairness all by themselves, but can be used as indicators of the decision-maker's underlying attitudes toward the costs of error, where it is these underlying attitudes that determine whether the decision-maker is *prima facie* fair or biased. (And, of course, even these attitudes do not yet determine *ultima facie* fairness, since there are cases in which caring differently about errors across groups is compatible with, or even required by, having equal regard for the interests of those groups.) Indeed, for black-box algorithms, any approach must be outcome-based in this epistemological sense – the algorithm's outputs are the only information we have on which to base our assessments of fairness, and so they must be used as indicators of whatever we take fairness to consist in simply because no other evidence is available.

In this project, our hope has been to start a conversation about the normative meaning of decision thresholds and their relationship to loss functions and priors. While this still is outcome-based in an epistemological sense, the way that outcomes enter into our picture is quite different from the way they enter into simple approaches that regard fairness as a function of the outcome distribution. In that sense, we hope to have made clear that our approach is not just a variation of an outcome based approach (such as Hellman (2020)'s, for example). Outcomes provide evidence of attitudes, where it is these attitudes in which biased or equal degrees of regard fundamentally lie. And observed outcomes provide evidence of these underlying attitudes by enabling us to estimate the decision-maker's assessment of the relative goodness or badness *of* the expected outcomes of her decisions for different groups (i.e. false positives, false negatives, true positives and true negatives). Given these complex interconnections between outcomes and attitudes, we hope that our project sheds further light on how traditional moral dichotomies – for example, between consequentialist and deontological perspectives or between outcome-based and procedural approaches – make it harder to understand the type of sophisticated moral evaluation in which we think we are required to engage in the context of algorithmic decision-making.

5 Concluding Remarks

In short, the threshold approach that we have articulated and defended here is the right approach to algorithmic fairness for those of us who recognize that whether a decision-maker holds individuals or groups in equal regard is a matter of the relative degrees to which she cares about what happens to them – a form of equal regard to whose absence we respond, rightly, with resentment and indignation.

References

Aigner, D. J. and G. G. Cain (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review 30*(2), 175–187.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). Machine Bias. *ProPublica* (May 23, 2016).

Arnold, C. (2016). Graduates Of Historically Black Colleges May Be Paying More For Loans: Watchdog Group. *NPR*.

Arpaly, N. (2000). On Acting Against One's Best Judgment. *Ethics 110*(3), 488–513.

Arpaly, N. and T. Schroeder (2013). *In Praise of Desire*. Oxford: Oxford University Press.

Arrow, K. J. (1972a). Models of Job Discrimination. In A. H. Pascal (Ed.), *Racial Discrimination in Economic Life*, pp. 83–102. Lexington, Mass.: Lexington Books, D. C. Heath and Co.,.

Arrow, K. J. (1972b). Some Mathematical Models of Race in the Labor Market. In A. H. Pascal (Ed.), *Racial Discrimination in Economic Life*, pp. 187–2042. Lexington, Mass.: Lexington Books, D. C. Heath and Co.,.

Arrow, K. J. (1974). The Theory of Discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*, pp. 1–33. Princeton University Press.

Autor, D. H. (2003). Lecture Note: The Economics of Discrimination.

Babic, B. (2019). A Theory of Epistemic Risk. *Philosophy of Science 86*(3), 522–550.

Babic, B., A. Gaba, I. Tsetlin, and R. L. Winkler (2021). Normativity, Epistemic Rationality, and Noisy Statistical Evidence. *British Journal for the Philosophy of Science* (Accepted April 30, 2021).

Babic, B., S. Gerke, T. Evgeniou, and I. G. Cohen (2021a). Beware Explanations from AI in Health Care. *Science 373*(6552).

Babic, B., S. Gerke, T. Evgeniou, and I. G. Cohen (2021b). Direct-to-Consumer Medical Machine Learning and Artificial Intelligence Applications. *Nature Machine Intelligence 3*, 283–287.

Babic, B., S. Gerke, T. Evgeniou, and I. G. Cohen (2021c). When Machine Learning Goes off the Rails. *Harvard Business Review* (January-February).

Basu, R. (2019). The Wrongs of Racist Beliefs. *Philosophical Studies 9*(176), 2497–2515.

Basu, R. and M. Schroeder (2019). Doxastic Wrongings. In B. Kim and M. McGrath (Eds.), *Pragmatic Encroachment in Epistemology*, pp. 181–205.

Becker, G. S. (1957). *The Economics of Discrimination* (1st ed.). Chicago: University of Chicago Press.

Benjamens, S., P. Dhunnoo, and B. Meskó (2020). The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database. *Nature Digital Medicine 3*.

Buchak, L. (2014). Belief, credence, and norms. *Philosophical Studies 169*(2), 285–311.

Cheng, E. K. (2013). Reconceptualizing the burden of proof. *Yale Law Journal 122*(5), 1254–1279.

Chohlas-Wood, A. (2020). Understanding Risk Assessment Instruments in Criminal Justice. *The Brookings Institution (Friday, June 19, 2020*.

Cohen, J. (1981). Subjective Probability and the Paradox of the Gatecrasher. *Arizona State Law Journal 1981*(2), 627–634.

Colyvan, M., H. M. Regan, and S. Ferson (2001). Is It a Crime to Belong to a Reference Class? *Journal of Political Philosophy 9*(2), 168–181.

Corbett-Davies, S. and S. Goel (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic Decision Making and the Cost of Fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*.

Flores, A., K. Bechtel, and C. Lowenkamp (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.". *Federal Probation 80*(2).

Foldessy, E. P. (1992). Largest Metropolitan Areas. *Wall Street Journal* (November 30, 1992, Table A8).

Furlough, C., T. Stokes, and J. Gillan, Douglas (2021). Attributing Blame to Robots: The Influence of Robot Autonomy. *Human Factors 63*(4), 592–602.

Gendler, T. S. (2011). On the Epistemic Costs of Implicit Bias. *Philosophical Studies* (1), 33–63.

Gerke, S., T. Minssen, and I. Cohen (2020). Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare. *Artificial Intelligence in Healthcare*, 295–336.

Gordon, C. (2021). The Rise Of AI In The Transportation And Logistics Industry. *Forbes* (September 5, 2021).

Hall, J. (2019). How Artificial Intelligence Is Transforming Digital Marketing. *Forbes* (August 21, 2019).

Hellman, D. (2020). Measuring Algorithmic Fairness. *Virginia Law Review 106*(4), 811–866.

Hidalgo, C., D. Orghian, J. Canals, F. Almeida, and N. Martin (2021). *How Humans Judge Machines*. Cambridge, MA: MIT Press.

Johnson, G. (2023). Are Algorithms Value Free. *Journal of Moral Philosophy* (Forthcoming).

Johnson King, Z. (2020). Don't Know, Don't Care? *Philosophical Studies 177*(2), 413–431.

Johnson King, Z. and B. Babic (2020). Moral Obligation and Epistemic Risk. In M. Timmons (Ed.), *Oxford Studies in Normative Ethics*, Volume 10, pp. 81–105.

Kaplan, J. (1968). Decision Theory and the Factfinding Process. *Stanford Law Review 20*(6), 1065–1092.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018). Algorithmic Fairness. *AEA Papers and Proceedings 108*, 22–27.

Kleinberg, J. and M. Mullainathan, S.and Raghavan (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.

Lander, E. and A. Nelson (2021). Americans Need a Bill of Rights for an AI-Powered World. *WIRED* (October 8, 2021).

Laplace, P. (1786). Sur les Naissances, les Mariages et les Morts à Paris Depuis 1771 Jusqu'à 1784 et Dans Toute L'étendue de la France, Pendant les Années 1781 et 1782. *Mémoires de l'Académie Royale des Sciences Présentés par Diverse Savans*.

Lima, G., N. Grgić-Hlača, and M. Cha (2021). Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems 372*, 1–17 (CHI'21).

Lima, G., N. Grgić-Hlača, and M. Cha (2023). Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems 372*, 1–26 (CHI'23).

Liptak, A. (2017). Sent to Prison by a Software Program's Secret Algorithms. *New York Times* (May 1, 2017).

McKenna, M. (2012). *Conversation and Responsibility*. Oxfor University Press.

Moss, S. (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.

Nabi, R., D. Malinsky, and I. Shpitser (2019). Learning Optimal Fair Policies. *Proceedings of the 36th International Conference on Machine Learning 97*, (ICML 36).

Nabi, R. and I. Shpitser (2018). Fair Inference on Outcomes. *Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence 32*(235), 1931–1940 (AAAI'18).

Nagel, T. (1976). Moral Luck. *Proceedings of the Aristotelian Society, Supplementary Volumes 50*, 137–155.

Nesson, C. (1985). The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts. *Harvard Law Review 98*(7), 1357–1392.

Phelps, E. S. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review 62*(4), 659–661.

Redmayne, M. (2008). Exploring the Proof Paradoxes. *Legal Theory 14*(4), 281–309.

Rimol, M. (2021). Gartner Forecasts Worldwide Artificial Intelligence Software Market to Reach $62 Billion in 2022. *Gartner* (November 22, 2021).

Robert, C. P. (2007). *The Bayesian Choice: From Decision Theoretic Foundations to Computational Implementation*. Springer.

Scanlon, T. (1998). *What We Owe to Each Other*. Belknap Press.

Schauer, F. (2003). *Profiles, Probabilities, and Stereotypes*. Cambridge: Harvard University Press.

Schwartz, O. (2019). Untold History of AI: Algorithmic Bias Was Born in the 1980s. *IEEE Spectrum*.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics 121*, 602–632.

Shoemaker, D. (2015). *Responsibility From the Margins*. Oxford: Oxford University Press.

Simoiu, C., S. Corbett-Davies, and S. Goel (2017). The Problem of Infra-Marginality in Outcome Tests For Discrimination. *The Annals of Applied Statistics 11*(3), 1193–1216.

Smith, A. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life. *Ethics 115*(2), 236–271.

Spence, A. M. (1973). Job Market Signaling. *Quarterly Journal of Economics 87*(3), 355–374.

Spence, A. M. (1974). *Market Signaling: Informational Transfer in Hiring and Related Screening Processes.* Cambridge: Harvard University Press.

Strawson, P. F. (1982). Freedom and Resentment. In G. Watson (Ed.), *Free Will* (1st ed.)., pp. 59–80. Oxford: Oxford University Press.

Thomson, J. J. (1986). Liability and individualized evidence. *Law & Contemporary Problems 49*(3), 199–219.

Tribe, L. H. (1971). Trial by Mathematics: Precision and Ritual in the Legal Process. *Harvard Law Review 84*(6), 1329–1393.

Veloso, M., T. Balch, D. Borrajo, P. Reddy, and S. Shah (2021). Artificial Intelligence Research in Finance: Discussion and Examples. *Oxford Review of Economic Policy 37*(3), 564–584.

Verma, S. and J. Rubin (2018). Fairness Definitions Explained. In Y. Brun, B. Johnson, and A. Meliou (Eds.), *Proceedings of the International Workshop on Software Fairness*, pp. 1–7. ACM.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics 24*, 227–248.

Watson, G. (2004). Responsibility and the Limits of Evil. In *Agency and Answerability: Selected Essays*, pp. 219–259. Oxford: Oxford University Press.

Williams, B. (1981). *Moral Luck*. Cambridge: Cambridge University Press.

Winling, L. C. and T. M. Michney (2021). The Roots of Redlining: Academic, Governmental, and Professional Networks in the Making of the New Deal Lending Regime. *Journal of American History 108*, 42–69.

Yong, E. (2018). A Popular Algorithm Is No Better at Predicting Crimes Than Random People. *The Atlantic* (January 17, 2018).