3

What Can Agents Reasonably Endorse?

In Chapter 2, we offered an account of autonomy that is both compatible with a broad range of views and ecumenical in that it incorporates important facets of competing views. The key features of our account are that autonomy demands both procedural independence (i.e., competence and authenticity) and substantive independence (i.e., social and relational conditions that nurture and support persons in acting according to their values as they see fit, without overweening conditions on acting in accord with those values). Our next task is to draw on that conception of autonomy to better understand and evaluate algorithmic decision systems. Among those that we will consider are ones we introduced in Chapter 1, including risk assessment algorithms such as COMPAS and K-12 teaching evaluation systems such as EVAAS.

How, though, do we get from an account of an important moral value such as autonomy to an evaluation of complex socio-technical systems? We will do that by offering a view of what it takes to respect autonomy and to respect persons in virtue of their autonomy, drawing on a number of different normative moral theories. Our argument will proceed as follows. We start with a description of another K-12 teacher evaluation case – this one from Washington, DC. We then consider several puzzles about the case. Next, we provide our account of respecting autonomy and what that means for individuals' moral claims. We will explain how that conception can help us understand the DC case, and we will offer a *general* account of the moral requirements of algorithmic systems.[1] Finally, we will explain how our view sheds light on our foundational cases (i.e., *Loomis*, *Wagner*, and *Houston*).

3.1 IMPACT: NOT AN ACRONYM

In 2007, Washington, DC sought to improve its public school system ("DC schools") by implementing an algorithmic teacher assessment tool, IMPACT, the aim of which is to identify and remove ineffective teachers. In 2010, teachers with

[1] This argument originated in Rubel, Castro, and Pham, "Algorithms, Agency, and Respect for Persons."

https://doi.org/10.1017/9781108895057.003 Published online by Cambridge University Press

IMPACT scores in approximately the bottom 2 percent were fired; in 2011, teachers with IMPACT scores in approximately the bottom 5 percent were fired.[2]

There is a plausible argument for DC schools using IMPACT. The algorithm uses complex, data-driven methods to find and eliminate inefficiencies, and it purports to do this in an objective manner. Its inputs are measurements of performance and its outputs are a function of those measurements. Whether teachers have, say, ingratiated themselves to administrators would carry little weight in the decision as to whether to fire them. Rather, it is (ostensibly) their effectiveness as teachers that grounds the decision. Using performance measures and diminishing the degree to which personal favor and disfavor affect evaluation could plausibly generate better educational outcomes.

Nonetheless, DC schools' use of IMPACT was problematic. This is in part because IMPACT's conclusions were epistemically flawed. A large portion of a teacher's score is based on VAM that seeks to isolate and quantify a teacher's individual contribution to student achievement on the basis of annual standardized tests.[3] However, VAMs are poorly suited for this measurement task.[4] DC teachers work in schools with a high proportion of low-income students. At the time IMPACT was implemented, even in the wealthiest of the city's eight wards (Ward 3) nearly a quarter of students were from low-income families, and in the poorest ward (Ward 8), 88 percent of students were from low income families.[5] As one commentary on IMPACT notes, low-income students face a number of challenges that influence their ability to learn:

> These schools' student bodies are full of kids dealing with the toxic stress of poverty, leaving many of them homeless, hungry, or sick due to limited access to quality healthcare. The students are more likely to have an incarcerated parent, to be deprived of fresh or healthy food, to have spotty or no internet access in their homes, or to live in housing where it is nearly impossible to find a quiet place to study.[6]

Given the challenges of their students, it is not surprising that fewer teachers in Ward 8 than Ward 3 are identified by IMPACT as "high performing."[7]

The effects of poverty are confounding variables that affect student performance on standardized tests. For this reason, we cannot expect VAMs – which use only annual test scores to assess a teacher's individual contribution to student achievement – to reliably find the signal of bad teaching through the noise of student poverty. Indeed, the American Statistical Association warns that studies on VAMs

[2] O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Turque, "More than 200 D.C. Teachers Fired."

[3] Isenberg and Hock, "Measuring School and Teacher Value Added in DC, 2011–2012 School Year"; see also Walsh and Dotter, "Longitudinal Analysis of the Effectiveness of DCPS Teachers."

[4] For extensive discussion, see Amrein-Beardsley, *Rethinking Value-Added Models in Education*.

[5] Quick, "The Unfair Effects of IMPACT on Teachers with the Toughest Jobs."

[6] Quick.

[7] Quick.

"find that teachers account for about 1% to 14% of the variability in test scores, and that the majority of opportunities for quality improvement are found in the system-level conditions."[8] The American Statistical Association also notes that "[VAMs] have large standard errors, even when calculated using several years of data. These large standard errors make rankings [of teachers] unstable, even under the best scenarios for modeling."[9]

So IMPACT suffers from an *epistemic* shortcoming. Is there also a *moral* problem? One possibility is that IMPACT poses a moral problem in that it harms teachers; it is harmful for teachers to lose their jobs, and IMPACT scores are the basis for that loss and harm. This, however, is not enough to conclude that there is moral wrong. Firing teachers can be justified (e.g., for cause), though it harms them, and use of IMPACT may create enough student benefit to risk some harm to teachers. Moreover, IMPACT is not obviously unfair; its epistemic flaws may be evenly distributed among teachers.

If there is something wrong about using IMPACT, what does that have to do with the epistemic problem, and what does it have to do with autonomy? We will argue that a teacher who is fired is *wronged* when that firing is based on a system that they could not reasonably endorse. We explain the general argument in the next section. We then apply that argument to IMPACT and our other polestar cases in the remainder of the chapter.

3.2 AUTONOMY, KANTIAN RESPECT, AND REASONABLE ENDORSEMENT

In Chapter 2, we explained that autonomy and self-governance involve (among other things) the capacity to develop one's own conception of value and sense of what matters, and the ability to realize those values by guiding one's actions and decisions according to one's sense of value. We explained the relationship of this conception to Kantian views.[10]

[8] American Statistical Association, "ASA Statement on Using Value-Added Models for Educational Assessment: Executive Summary," 2; see also, Morganstein and Wasserstein, "ASA Statement on Value-Added Models."

[9] American Statistical Association, 7; see also, Morganstein and Wasserstein.

[10] It is worth reiterating several points from Chapter 2 to emphasize the limits of this Kantian formulation. The capacity to self-govern, the values agents develop, and the ways in which they incorporate those values into their lives are socially situated. See Mackenzie and Stoljar, *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, 4. Developing one's sense of what is important depends on social conditions that nurture the ability to do so. See Oshana, *Personal Autonomy in Society*, 90. Social structures may delimit the conceptions of value that are available for persons to draw upon in developing their sense of value. Persons' abilities to incorporate their values into their important decisions will depend on what opportunities exist in the broader social context. See Raz, *The Morality of Freedom*; Mackenzie, "Relational Autonomy, Normative Authority and Perfectionism." The fact that self-governance is socially situated, however, does not undermine the importance of autonomy and agency. Rather, failures to nurture persons' abilities to develop their

The issue we are addressing here, though, is what kinds of moral requirements are grounded in autonomy. How, in other words, does autonomy ground persons' moral claims? There are a number of different ways to address this question.

Let's begin with a prominent account by Christine Korsgaard.[11] The basic idea of autonomy – that is, that each of us in our capacity as autonomous beings develop conceptions of value for ourselves and act on those conceptions – is that people are self-legislators. By engaging in self-legislation, we understand our capacity to determine what matters for ourselves as a source of value. If we treat this capacity as a source of value, then it is the capacity itself (not, say, our own egoism) that must be valuable. Hence, any instances of that capacity (not just our own) must also be a source of value. So, because we are autonomous, we must value (which is to say, respect) autonomy *generally*. In other words, the premise that the capacity to self-legislate grounds value in one's own case entails a conclusion that a similar capacity to self-legislate must also ground value in others' cases.

A different way to ground the value of autonomy is its connection to well-being. Individuals have the capacity to develop their own sense of value; they are generally well positioned to understand how to advance that value, and the ability to do so (within reasonable parameters) is an important facet of their well-being. Because we have good reasons to promote well-being in ourselves and others, we therefore have good reasons to respect autonomy in ourselves and others. This is the line of reasoning that a utilitarian, for example, John Stuart Mill, can use in support of respecting autonomy.[12]

Views like these link the concept of autonomy to the moral value of respecting autonomy. But what does respect for autonomy require? Returning to Kant, there are different but (roughly) equivalent ways to spell this out. One way to respect autonomy is to abide by the second, Humanity Formulation of the Categorical Imperative:

> **Humanity Formulation:** So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means.[13]

Treating something as an end requires treating it as something that is valuable in its own right; treating something "merely" as a means involves treating it as solely an instrument for the promotion of an end, without also treating it as an end itself.

agency and substantial constraints on options available for incorporating values into persons' lives are moral problems in part because of the importance of autonomy. See, for instance, Superson, "Deformed Desires and Informed Desire Tests"; Meyers, "Personal Autonomy and the Paradox of Feminine Socialization." Hence, even though our conception of autonomy echoes Kantian views, it would be a mistake to conclude that autonomy in this sense assumes that individuals are separable from their social, familial, and relational lives.

[11] Korsgaard et al., *The Sources of Normativity*.

[12] Mill, *On Liberty*, chapter 3.

[13] Kant, *Groundwork of the Metaphysics of Morals*, sec. 4:429.

Treating someone as an end-in-themselves requires that we take seriously their ability to make sense of the world and their place in it, to determine what matters to them, and to act according to their own understanding and values (to the extent that they see fit). They may not be considered solely in terms of how they advance values of others.

The Humanity Formula is, of course, vague and has long been the subject of dispute. Derek Parfit can provide some help in specifying it. He offers the following principle as the core idea of the Humanity Formula:

> **Consent Principle:** It is wrong to treat anyone in any way to which this person could not rationally consent.[14]

Famously, Kant gave several formulations of the Categorical Imperative, of which the Humanity Formula is just one. According to another formulation,

> **Formula of Universal Law:** Act only in accordance with that maxim through which you can at the same time will that it become a universal law.[15]

A "maxim" is a principle that connects an act to the reasons for its performance. Suppose one makes a donation to Oxfam. Their maxim might be, "In order to help reduce world hunger, I will contribute fifty dollars a month to Oxfam." An act is morally permissible when its maxim is universalizable, that is, if (and only if) every rational person can consistently act on it.

The Formula of Universal Law is often compared to the Golden Rule. This is a comparison Kant would be loath to accept since he rejected the Golden Rule as a moral principle. Parfit, in his inimitable style, thinks that "Kant's contempt for the Golden Rule is not justified."[16] And indeed, Parfit offers a reconstruction of the Golden Rule that incorporates the core ideas of the Formula of Universal Law as follows:

> **Golden Rule:** We ought to treat *everyone* as we would rationally be willing to be treated if we were going to be in all of these people's positions, and would be relevantly like them.[17]

As consideration of the Humanity Formula, the Consent Principle, the Formula of Universal Law, and the Golden Rule help to lay bare, respecting autonomy involves both an element of treating others in ways to which they can agree (because it aligns with their ends, for example) and an element of understanding how others' positions are relevant to which ends one adopts as one's own.

Hence, there are two facets to this "golden rule" style formulation. The first has to do with a person's ability to develop and endorse their sense of value and act

[14] Parfit, *On What Matters*, 181.
[15] Kant, 4:421.
[16] Parfit, 19.
[17] Parfit, 327.

accordingly, including what treatment would they willingly subject themselves to. This facet of the rule explains wrongs that are associated with deception. Deception is wrong (when it is) in part because it circumvents an agent's ability to make decisions according to their own reasons. Likewise, paternalism is an affront (when it is) because it supplants a person's ability to act on their own reasons based on a degree of distrust of their agency.[18]

The second facet of the Golden Rule has to do with the treatment of others. Autonomy can underwrite moral claims only to the extent that it is used to ends that are compatible with others' reasonable interests. The requirement that we consider how we would be rationally willing to be treated if we were relevantly similar to, and in similar circumstances as others is a way of making vivid others' reasonable interests. It also echoes Joel Feinberg's understanding of "autonomy as ideal" (as we discussed in Section 2.2.2). Autonomy as ideal recognizes that people can exercise autonomy badly (such that facets of autonomy are not necessarily virtues) and that people are parts of larger communities. Hence, Feinberg explains, the ideal of an autonomous person requires that their self-governance be consistent with the autonomy of others in their community.[19] This, in turn, reflects Kant's understanding that morally right action requires that the action can coexist with everyone else's ability to exercise freedom under universal moral law.[20]

Feinberg's understanding of autonomy as ideal is reflected in two other conceptions of respecting autonomy that are useful in developing our view. The first comes from John Rawls. In developing his understanding of just political and social systems, Rawls describes people as having two moral powers. The idea is that any person in the original position – which is to say anyone deciding on the basic structure of the society in which they will live, but knowing nothing of their place in it and nothing about their particular characteristics – must possess two powers for their choices to make sense. First, they must be rational. As in our discussion in Chapter 2, "rational" here just means the ability to engage in basic reasoning about means and ends, coupled with some set of basic values and motivations. The idea is that for a person to prefer one social and political structure over another, they must have some basic motivations to ground that preference. If literally nothing mattered to an individual, there would be no basis for their choices. Second, persons must be reasonable. This simply means that they are willing to abide fair terms of social cooperation, so long as others do too. It requires neither subordinating one's reasonable interests to others nor accepting outlandish demands from others.

Rawls's view is that people with these two powers in the original position would, for reasons having to do with nothing more than their own self-interest, accept certain social structures as binding. They will advance ends that people endorse (after all, those ends might be their own) and will establish fair terms of social

[18] Shiffrin, "Paternalism, Unconscionability Doctrine, and Accommodation."
[19] Feinberg, "Autonomy," 44–45.
[20] Kant, *Groundwork of the Metaphysics of Morals*, sec. 6:230.

cooperation because they will be in a position where they will have to abide those terms. Now, there are myriad criticisms and limitations of Rawls's view, but his conception is useful in that it connects procedural autonomy (or psychological autonomy, as we described it in Chapter 2) to respect and social cooperation. Following Kant, Rawls's view is that persons' exercise of their own autonomy is important, but justifiable only to the extent that it is compatible with others'. And, hence, principles *limiting* autonomy can be grounded in fair terms of social cooperation.

A different view comes from Scanlon. Both Scanlon and Rawls are grounded in social contract theories. However, Rawls's target is society's basic structure while Scanlon's main concern is to articulate basic moral principles governing social interaction. Moreover, while Rawls derives principles based on people rationally advancing their own self-interest, Scanlon aims to derive principles based on an account of the reasons one can offer to others to justify conduct. Specifically, Scanlon argues that "[a]n act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced, general agreement."[21] Parfit distills this view into what we might call the "reasonable rejection" criterion: "Everyone ought to follow the principles that no one could reasonably reject."[22] This criterion holds the linchpin of morality to be the strength of people's reasons: If one has good reasons against some principle and actions based on it, but others have weightier reasons *for* that principle and actions based on it, those weightier reasons should prevail on the grounds that one cannot reasonably reject them.

Rahul Kumar characterizes Scanlon's contractualism as grounding persons' legitimate expectations and demands of one another concerning conduct and consideration "as a matter of basic mutual respect for one another's value as rational self-governors."[23] A key facet of Scanlon's approach, and one that unites the respect-for-persons views we are drawing on here, is that each requires paying attention to individuals and to the separateness of persons.[24] To understand this, contrast the requirement that actions be based on principles no one could reasonably reject with aggregating views such as some forms of consequentialism (i.e., those that are concerned with aggregated welfare).

As Kumar notes, consequentialist concerns with our actions are subordinate to (i.e., only matter in light of) the results of those actions. However, contractualism (and autonomy-respecting theories *generally*) focuses on how our actions reflect our relationships with others directly. Consequences, on this view, matter only insofar as they reflect respect for other persons. That is,

> [Consequentialism is] concerned with what we do, but only because what we do affects what happens. The primary concern in [consequentialism] is with the promotion of well-being. Contractualism is concerned with what we do in a more basic

[21] Scanlon, *What We Owe to Each Other*, 153.
[22] Parfit, *On What Matters*, 360.
[23] Kumar, "Reasonable Reasons in Contractualist Moral Argument," 10.
[24] Rawls, *A Theory of Justice*, sec. 5.

https://doi.org/10.1017/9781108895057.003 Published online by Cambridge University Press

> sense, since the reasons for which we act express an attitude toward others, where what is of concern is that our actions express an attitude of respect for others as persons.[25]

Respecting others as having their own sense of value and being able to order their lives accordingly also makes it the case that their expectations should matter to us, and mutual respect entails that people have legitimate expectations for what they can expect of others. People have good reason to expect that others will respect them as having their own ends and as being capable of abiding fair terms of social cooperation. Treatment that frustrates that expectation is a failure of respect. As Kumar puts it,

> Disappointments of such expectation are (at least *prima facie*) valid grounds for various appropriate reactive attitudes toward one another. Resentment, moral indignation, forgiveness, betrayal, gratitude – the range and subtlety of reactions we have toward others with whom we are involved in some kind of interpersonal relationship is inexhaustible – all presuppose beliefs about what we can reasonably expect from others.[26]

We will return to the issue of reactive attitudes in Chapter 7, where we examine automated decision systems and responsibility.

So what is the upshot?

Recall that the purpose of this section is to move from an understanding of autonomy (from Chapter 2) and explain some moral claims that are grounded in respecting autonomy. We have drawn on several views about respecting autonomy, each of which attends to the importance of the principles one wills for oneself and to the incontrovertible fact that humans are social beings, and hence, to the fact that human moral principles require broad and deep social cooperation. Respecting autonomy, in other words, requires both attention to individuals' conceptions of their own good *and* some broad conception of social cooperation. Notice, though, that the views we have drawn on have substantial overlap. They may entail some differences in application (though that is an open question and would merit an argument), and they may have slightly different normative grounds for principles. But our project is deliberately ecumenical, and for our purposes the most important thing is the similarity across these views. They all point in the same direction, and they each provide a foundation on which to articulate the following, which captures their key elements, at least to a first approximation.

> **Reasonable Endorsement Test:** An action is morally permissible only if it would be allowed by principles that each person subject to it could reasonably endorse.

According to this test, then, subjecting a person to an algorithmic decision system is morally permissible only if it would be allowed by principles that everyone could reasonably endorse.

It's worth clarifying a couple of points about the Reasonable Endorsement Test. It differs from the articulations given by Scanlon and Rawls. One reason that Scanlon uses

[25] Kumar, "Defending the Moral Moderate: Contractualism and Common Sense," 285.
[26] Kumar, 286.

https://doi.org/10.1017/9781108895057.003 Published online by Cambridge University Press

"reasonably reject" is to emphasize that persons must compare the burdens that they must endure under some state of affairs with others' burdens under that state of affairs. Hence, if a person would reject a social arrangement as burdensome, but their burden is less than others, there are substantial benefits to the arrangement, the alternatives are at least as burdensome to others, and the overall consequences are not substantially better, then the person's rejection of the arrangement would be unreasonable.

Our use of "could reasonably endorse" does similar work, as we make clear in our discussion of IMPACT later. However, by focusing on endorsement it leans closer to Parfit's reformulation of the Golden Rule. Specifically, people can reasonably endorse (i.e., can be rationally willing to be treated according to) principles as either consistent with their own sense of value *or* as fair terms of social cooperation. Note, too, that actions and principles that do not affect an individual's personal interests are nonetheless candidates for reasonable endorsement because those individuals can evaluate them as fair terms of social cooperation.

Another thing to note is that each of the formulations we have drawn on, as well as our Reasonable Endorsement Test, will inevitably have important limitations. Recall that each is trying to provide a framework for guiding actions based on a more basic moral value (autonomy). Hence, what matters for each principle is the extent to which it recognizes and respects persons' autonomy. For the reasons we outlined earlier, we think that the reasonable endorsement principle does at least as well in capturing respect for autonomy as the other formulations.

Finally, even if another principle does a better job reflecting the nature and value of autonomy and providing guidance, our sense is that those views, when applied, will converge with ours (at least in the cases that are of interest here). And in any case, objecting to our larger project on the grounds that a different kind of social agreement principle better captures autonomy and social cooperation warrants an argument for why it would better explain concerns in the context of algorithmic systems.

3.3 TEACHERS, VAMS, AND REASONABLE ENDORSEMENT

To sum up our view so far: teachers are autonomous persons, and hence they have a claim to incorporate their values into their lives as they see fit. And *respecting* them requires recognizing them as value-determiners, neither thwarting nor circumventing their ability to act according to those values without good reason. They are also capable of abiding fair terms of social agreement (so long as others do too), and hence "good reasons" for them will be reasons they can endorse as fair terms of social cooperation, which means they can endorse those reasons as either consistent with their own values or as a manifestation of fair social agreement.

Now, what is it to thwart an agent's ability to act according to their values? One example, discussed earlier, is deceit, in which one precludes an agent's ability to understand circumstances relevant to their actions. Another way to thwart agency is to create conditions in which agents are not treated according to reasons that they

could reasonably endorse, were they given the opportunity to choose how to be treated. That is, precluding persons from acting according to their values (e.g., by deceit) or placing them in circumstances that they cannot endorse as fair is a failure of recognition of them as value-determiners and a form of disrespect.

IMPACT fails to respect teachers in exactly this way (i.e., placing them in circumstances they cannot endorse), for several interrelated reasons.[27] The reasons are reliability, responsibility, stakes, and relative burden, and they work as general criteria for when people can reasonably endorse algorithmic systems.

Reliability. For the purposes of this project, we will understand reliability in its colloquial sense; that is, as consistent (though not necessarily infallible) accuracy.[28] We have provided some reasons for why IMPACT is an unreliable tool for the evaluation of teacher efficacy. Now, teachers, like any professionals, can reasonably endorse a system in which they are evaluated based on their efficacy. Moreover, through their training and professionalization, they have endorsed the value of educating students, and fair terms of social cooperation would require that truly ineffective teachers be identified for this reason. But because IMPACT is unreliable, there is some reason to think that it misidentifies teachers as ineffective. Hence, teachers should be loath to endorse being evaluated by IMPACT.

Responsibility. IMPACT's lack of reliability is not the only way it fails to respect autonomy. Imagine a case where a teacher evaluation system reliably measures student learning. Two teachers score poorly in this year's assessment. One scores poorly because she did not assign curriculum-appropriate activities, while the other scores poorly because her classroom lacks air-conditioning. Only the first teacher is responsible for her poor scores. The second teacher's scores are based on factors for which she is not responsible. Teachers could not reasonably endorse such a system.[29]

Given the population many DC teachers were working with – underserved students – IMPACT cannot be understood as tracking only factors for which teachers are responsible. The effects of poverty, abuse, bullying, illness, undiagnosed

[27] To be clear, we think that each of these dimensions is relevant in determining whether use of an algorithm is morally problematic. However, we do not think that the dimensions we outline are exhaustive; this list is not meant to be comprehensive. There may be other considerations, such as consideration of desert or other facets of fairness which can play an important role in assessing the appropriateness of the use of an algorithm.

[28] Because we are using "reliability" in its colloquial sense, we will *not* be using the term in the statistical (and more captious) sense of being free from random error; our use of reliability will more closely align with the statistical sense of "validity," that is, accuracy borne from use of a (statistically) reliable method. We refrain from using the term "validity" to avoid confusion with the philosophical sense of the term, that is, premises entailing the conclusion of an argument. For more on the statistical senses of reliability and validity and an appraisal of value-added models in those terms, see Amrein-Beardsley, *Rethinking Value-Added Models in Education*, chapter 6.

[29] Notice that in this example responsibility and reliability are both relevant. Teachers could reasonably endorse a system in which their jobs depend on factors for which they are not responsible – e.g., population decline. However, firing teachers whose scores suffer because of exogenous factors (lack of air-conditioning) involves criteria that are not teachers' responsibilities and which are unreliable in making teaching better (though perhaps reliable in achieving better learning outcomes).

learning disabilities (resources for addressing these are much more limited in underserved districts), and so on plausibly undermine teacher efficacy. Yet teachers bear no responsibility for those impediments. So, even if the VAMs were reliable, teachers could not reasonably endorse their implementation.

Note that the dimension of responsibility not only covers the factors that teachers *can't* be responsible for (e.g., children's circumstances outside of school), but also factors they *shouldn't* be held responsible for. It is not impossible to imagine, for instance, that teachers who bring snacks to every session could motivate their students to get higher test scores. Or, that teachers who repair air-conditioners themselves do. Teachers who do (or do not) bring snacks, fix air-conditioners, etc., *can* be responsible for engaging in (or refraining from) those activities, in the sense that they have the power to engage in (or refrain from) those activities. However, they *shouldn't* be held responsible for refraining from the activities mentioned here, as this is not a reasonable ask.

Now, what exactly makes for a reasonable ask? That is a question that we cannot give an informative general account of, as it will vary greatly from domain to domain. We simply introduce the dimension of responsibility to our criterion to highlight the fact that algorithmic systems can affect persons for factors they either cannot or should not be responsible for, and that one factor relevant to the question of whether someone *may* be affected for partaking in (or refraining from) an activity is whether asking them to partake in (or refrain from) that activity is itself reasonable.

Stakes. Perhaps the most important factor in determining whether agents can reasonably endorse an algorithmic decision system is the stakes involved. Suppose that a VAM is set up to provide teachers with lots of information about their own practices but is not used for comparative assessment. The scores are shared with teachers privately and are not used for promotion and firing. Such a system might not be very reliable, or it might measure factors for which teachers are not responsible. Nonetheless, teachers might endorse it despite its limitations because the stakes are low. But if the stakes are higher (work assignments, bonuses, promotions), it is reasonable for the employees to want the system to track factors which can be reliably measured and for which they are responsible.

DC schools' use of IMPACT is high stakes. Teachers rely on their teaching for a paycheck, and many take pride in what they do. They have sought substantial training and often regard educating students as key to their identities. Having a low IMPACT score might cost a teacher their job and career, and it may well undermine their self-worth. By agreeing to work in particular settings they have formed reasonable expectations that they can continue to incorporate those values into their lives, subject to fair terms of cooperation (e.g., that they do their work responsibly and well, that demand for their services continues, that funding remains available, etc.).

IMPACT does poorly on our analysis. It is not reliable, it evaluates teachers based on factors for which they are not responsible, and it is used for high-stakes decisions. These points are reflected in teacher reactions to IMPACT. For example, Alyson

Perschke – a fourth-grade teacher in DC schools – alleged in a letter to Chancellor Kaya Henderson that VAMs are "unreliable and insubstantial."[30] Perschke did so well in her in-class observations that her administrators and evaluators asked if she could be videotaped as "an exemplar."[31] Yet the same year her VAM dragged her otherwise-flawless overall evaluation down to average. Remarking on this, she says, "I am baffled how I teach every day with talent, commitment, and vigor to surpass the standards set for me, yet this is not reflected in my final IMPACT score."[32]

Relative Burden. Another factor that is relevant in determining whether persons subject to an algorithmic system can reasonably endorse it is what we will call "relative burden." It is plausible that IMPACT disproportionately negatively affects teachers from underrepresented groups:

> [T]he scarcely mentioned, uglier impact of IMPACT is disproportionately felt by teachers in DC's poorest wards – at schools toward which minority teachers tend to gravitate. In seeking to improve the quality of teachers, IMPACT manages to simultaneously perpetuate stubborn workforce inequalities and exacerbate an already alarming shortage of teachers of color.[33]

So IMPACT might impose more burdens on members of underrepresented groups. This is another reason – independent of reasons grounded in reliability, responsibility, and stakes – that teachers should refrain from endorsing IMPACT.

There has been a great deal of discussion about how algorithmic systems may be *biased* or *unfair*. However, precisely what those concepts amount to and the degree to which they create moral problems are often unclear. Indeed, different conceptions of fairness can lead to different conclusions about whether a particular system is unfair (cf. Section 3.5).[34]

The problem, stated generally, is that there are many ways in which a decision system can represent subpopulations differently. The data on which a system is built and trained may under- or overrepresent a group. The system may make predictions

[30] Strauss, "D.C. Teacher Tells Chancellor Why IMPACT Evaluation Is Unfair."

[31] Strauss.

[32] Strauss. There is another autonomy-related issue here. In Chapter 2, we explained the importance of social and relational facets of autonomy. One way to understand the relationship between autonomy and facts about persons' social circumstances and relationships is that social and relational facts are *causally* important in fostering persons' autonomy. Another way is to understand social and relational facts as *constitutive* of autonomy. See Mackenzie and Stoljar, *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, chapter 1. That is, being a part of supportive and meaningful social groups and relationships is (1) a necessary condition for developing the competences and authenticity necessary for psychological autonomy and (2) an inherent part of being a person incorporating values into their life. Teacher-student and teacher-community relationships are deeply important and constitutive of the lives that many teachers value and have cultivated so as to realize their sense of value. Subjecting those relationships to unreliable, high-stakes processes that measure things for which teachers are not responsible conflicts with *that* facet of autonomy as well. Note that this is distinct from the reasonable endorsement argument.

[33] Quick, "The Unfair Effects of IMPACT on Teachers with the Toughest Jobs."

[34] Corbett-Davies and Goel, "The Measure and Mismeasure of Fairness."

about facts that have different incidence rates in populations, leading to different false-positive and false-negative rates. It is plausible to describe each of these cases of difference as in some sense "unfair." But we cannot know as a general matter *which conception* of fairness is appropriately applied across all cases. Thus, we will need an argument for applying any particular conception. In other words, even once we have determined that an algorithmic system is in some sense unfair, there is a further question as to whether (and *why*) it is morally problematic.

Put another way, one could frame our argument about the conditions under which agents can reasonably endorse algorithmic systems as an argument about the conditions under which such systems *are fair* (if everyone could reasonably endorse a system, that system is ipso facto fair). On that framing, fairness is the conclusion of our analysis, so including fairness as a general criterion for whether agents could reasonably endorse a system would make our argument circular.[35] Simply pointing out differences in treatment and concluding that those differences are unfair is not enough to move our argument forward. Rather, our task here is to identify factors that matter in determining whether people can reasonably endorse a decision system independently of whether they can be characterized as unfair. Among those is relative burden.

All automated systems will distribute benefits and burdens in some way or another.[36] Smart recommender systems for music, for example, will favor some artists over others and some users' musical tastes over others. Of course, what matters for our view here is whether they can reasonably endorse such a system. Relative burden matters if that burden either (a) is arbitrary, such that it has nothing to do with the system in the first place, (b) reflects otherwise morally unjustifiable distinctions, or (c) is a *compound* burden, which reflects or exacerbates an existing burden on a group.

An example of an arbitrary burden (a) would be a test that systematically scored English teachers more poorly overall than teachers of other subjects and thereby created some social or professional consequences for those teachers. An example of (b) would be a test that, let's suppose, systematically scored kindergarten teachers who are men lower than others. Perhaps kindergarteners respond less well to even very well-qualified men than they do to women; it would be morally unjustifiable, though, to have evaluation systems reflect that (at least if the stakes are significant). The unjustifiability in that case, though, is not related to some other significant social disadvantage. An example of (c) would be a case where an automated system imposes a burden that correlates with some other significant social burden (in many cases race, ethnicity, gender, or socioeconomic position).

[35] To be clear, we are not arguing that fairness analyses are mistaken. Rather, there are many conceptions of fairness that focus on different values. These may conflict, and many are mutually incompatible. Hence, there is a great deal of work to do in working out fairness issues even once one determines that a system is in some sense fair or unfair. We take this issue up again in Section 3.5.

[36] Note that the relationships between system burden and social circumstances need not be causal.

3.4 APPLYING THE REASONABLE ENDORSEMENT TEST

So far, we have argued that for an algorithmic system to respect autonomy in the relevant way, those who are subject to the system must be able to reasonably endorse it. And whether people can reasonably endorse a given system will be a function of its reliability, the extent to which it measures outcomes for which its subjects are responsible, the stakes at hand, and the relative burden imposed by the system's use. We illustrated our framework by analyzing DC Schools' use of IMPACT. Here, we turn back to our polestar cases to help make sense of the moral issues underlying them.

3.4.1 Wagner v. Haslam

Recall from Chapter 1 that the plaintiffs in this case (Teresa Wagner and Jennifer Braeuner) were teachers who challenged the Tennessee Value-Added Assessment System (TVAAS), which is a proprietary system similar to IMPACT. Because TVAAS did not test the subjects Wagner and Braeuner taught, they were evaluated based on a school-wide composite score, combined with their (excellent) scores from in-person teaching observations. This composite score dragged their individual evaluations from the highest possible score (as they had received in previous years) to middling scores. As a result, Wagner did not receive a performance bonus and Braeuner was ineligible for consideration for tenure. Moreover, each "suffered harm to her professional reputation, and experienced diminished morale and emotional distress."[37]

There is a deeper moral issue grounding the legal case. Wagner and Braeuner frame their case in terms of harms (losing a bonus, precluding tenure consideration, and so forth), but those harms matter only because they are wrongful. They are wrongful because TVAAS is an evaluation system that teachers could not reasonably endorse. Wagner and Braeuner's scores did not reliably track their performances nor did the scores reflect factors for which they were responsible, as the scores were based on the performance in subjects Wagner and Braeuner did not teach. And the stakes in the case are fairly high (there were financial repercussions for Wagner and job security for Braeuner). So, per our account, they were wronged.

There may also be a relative burden issue with TVAAS, though it is not discussed explicitly in the *Wagner* opinion. A 2015 study of TVAAS found that mathematics teachers across Tennessee were, overall, found to be more effective by TVAAS than their colleagues in English/language arts.[38] This finding supports two hypotheses: Tennessee's math teachers are more effective than its English/language arts teachers, and TVAAS is systematically biased in favor of math teachers.[39] If the

[37] *Wagner v. Haslam*, 112 F. Supp. 3d at 690.
[38] Holloway, "Evidence of Grade and Subject-Level Bias in Value-Added Measures."
[39] Amrein-Beardsley, "Evidence of Grade and Subject-Level Bias in Value-Added Measures: Article Published in TCR"; Spears, "Bias Confirmed – Tennessee Education Report."

latter hypothesis is true – and we suspect that it is – then, there are teachers, specifically Tennessee's English/language arts teachers, who have an additional complaint against TVAAS: It imposes a higher relative burden on them because it arbitrarily returns lower scores for non-math teachers. They could not reasonably endorse this arrangement.

3.4.2 Houston Fed of Teachers v. Houston Ind Sch Dist

The *Houston Schools* case is superficially similar to *Wagner* in that it involves a similar proprietary VAM (EVAAS) to evaluate teachers. The school system used EVAAS scores as the sole basis for "exiting" teachers.[40] The primary concern for our purposes is that Houston Schools did not have a mechanism for ensuring against basic coding and clerical errors. They refused to correct errors on the grounds that doing so would require them to rerun their analysis of the entire school district. That, in turn, would have two consequences. First, it would be costly; second, it would "change all other teachers' reports."[41]

The moral foundations of the teachers' complaints should by now be clear. The stakes here – i.e., losing one's job and having one's professional image tarnished – are high. EVAAS is unreliable, having what the court called a "house-of-cards fragility."[42] And that unreliability is due to factors for which teachers are not responsible, "ranging from data-entry mistakes to glitches in the code itself."[43] Hence, teachers could not reasonably endorse being evaluated under such a system.

We can add a complaint about relative burden, at least for some teachers. EVAAS, like IMPACT, gives lower scores to teachers working in poorer schools. To see this, consider an analysis of EVAAS's use in Ohio (in 2011–12) conducted by The Plain Dealer and State Impact Ohio, which found the following:

- Value-added scores were 2½ times higher on average for districts where the median family income is above $35,000 than for districts with income below that amount.
- For low-poverty school districts, two-thirds had positive value-added scores – scores indicating students made more than a year's worth of progress.
- For high-poverty school districts, two-thirds had negative value-added scores – scores indicating that students made less than a year's progress.
- Almost 40 percent of low-poverty schools scored "Above" the state's value-added target, compared with 20 percent of high-poverty schools.
- At the same time, 25 percent of high-poverty schools scored "Below" state value-added targets while low-poverty schools were half as likely to score "Below."[44]

[40] *Houston Fed of Teachers, Local 2415 v. Houston Ind Sch Dist*, 251 F. Supp. 3d at 1175.
[41] Houston Independent School District, "EVAAS/Value-Added Frequently Asked Questions."
[42] *Houston Fed of Teachers, Local 2415 v. Houston Ind Sch Dist*, 251 F. Supp. 3d at 1178.
[43] *Houston Fed of Teachers, Local 2415 v. Houston Ind Sch Dist*, 251 F. Supp. 3d at 1177.
[44] Ideastream, "Grading the Teachers."

In virtue of these findings, it is plausible that – in addition to complaints grounded in reliability, responsibility, and stakes – teachers in low-income schools have a complaint grounded in the uneven distribution of burdens.

And, hence, use of EVAAS is not something teachers subject to it could reasonably endorse.

3.4.3 Wisconsin v. Loomis

Our framework for understanding algorithmic systems and autonomy applies equally well to risk assessment tools like COMPAS.

To begin, COMPAS is moderately reliable. Researchers associated with Northpointe assessed COMPAS as being accurate in about 68 percent of cases.[45] More important is that COMPAS incorporates numerous factors for which defendants are not responsible. Recall that among the data points that COMPAS takes into account in generating risk scores are prior arrests, residential stability, employment status, community ties, substance abuse, criminal associates, history of violence, problems in job or educational settings, and age at first arrest.[46] Regardless of how well COMPAS's big and little bars reliably reflect reoffense risk, defendants are not responsible for some of the factors that affect those bars. So, while Loomis did commit the underlying conduct and was convicted of prior crimes, COMPAS incorporates factors for which defendants are not responsible.[47] For example, the questionnaire asks about the age at which one's parents separated (if they did); whether one was raised by biological, adoptive, or foster parents; whether a parent or sibling was ever arrested, jailed, or imprisoned; whether a parent or parent-figure ever had a drug or alcohol problem; and whether one's neighborhood friends or family have been crime victims.[48] Moreover, even if some factors (e.g., residential

[45] Brennan, Dieterich, and Ehret, "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System." We should note that how we apply the concept of reliability could itself be a matter of dispute. The study by Northpointe-affiliated researchers considers how well calibrated COMPAS is; that is, how likely COMPAS is to predict individual defendants' reoffense. However, there are other relevant measures for which COMPAS could be more or less reliable. A study by ProPublica found that prediction failure was different for White and Black defendants such that White defendants labeled lower risk were more likely to reoffend than Black defendants with a similar label, and Black defendants labeled higher risk were less likely to reoffend than White defendants labeled higher risk. See Angwin et al., "Machine Bias," May 23, 2016. These results call into question COMPAS's reliability in avoiding false positives and false negatives. We address this issue in more detail in the next section.

[46] Northpointe, Inc., "Practitioner's Guide to COMPAS Core," 24.

[47] Drawing on factors for which one is not responsible is compatible with a range of theories of punishment. Such factors may help determine a sentence – whether one is even arrested, moral luck, how well punishment deters crime, and so forth. But our view is not that only factors for which one is responsible may contribute to sentencing decisions. Rather, our view is that, as such factors increase, it becomes more difficult for an agent to abide such a system.

[48] Other questions pertain to matters for which defendants' responsibility is less clear: how often one has had barely enough money to get by, whether one's friends use drugs, how often one has moved in the last year, and whether one has ever been suspended from school.

stability, employment status, and community ties) are things over which individuals *can* exercise a degree of control, they are matters for which we *shouldn't* attribute responsibility in determining sentencing for offenses.

Further, the use of COMPAS in *Loomis* is high stakes. Incarceration is the harshest form of punishment that the state of Wisconsin can impose. This is made vivid by comparing the use of COMPAS in *Loomis* with its specified purposes. COMPAS is built to be applied to decisions about the type of institution in which an offender will serve a sentence (e.g., lower or higher security), the degree of supervision (e.g., from probation officers or social workers), and what systems and resources are appropriate (e.g., drug and alcohol treatment, housing, and so forth). Indeed, Northpointe warns against using COMPAS for sentencing, and Loomis's presentence investigation report specifically stated the COMPAS report should be used "to identify offenders who could *benefit from interventions and to target risk factors that should be addressed during supervision.*"[49] When the system is used for its intended purposes – identifying ways to mitigate risk of reoffense of persons under state supervision – the stakes are much lower.[50] Hence, it is more plausible that someone subject to its use could reasonably endorse it in those cases.

One of Loomis's complaints about COMPAS is that it took his gender into account. The court found that COMPAS's use of gender was not discriminatory because it served the purpose of promoting accuracy. So Loomis's claim that he shouldered a higher relative burden under this system was undercut, and the court was – in our opinion – correct in their response to his claim. Because men do commit certain crimes more often than women, removing gender as a factor could result in the systematic overestimation of women's risk scores.[51] This does not, however, mean that COMPAS has no issues with respect to the question of relative burden.

To introduce the relative burden issue, let's turn to a related controversy surrounding COMPAS. In May 2016, ProPublica reported that COMPAS was biased against Black defendants.[52] Specifically, ProPublica found that COMPAS misidentified Black defendants as high risk twice as often as it did White defendants. Northpointe, the company that developed COMPAS, released a technical report

[49] *Wisconsin v. Loomis*, 881 N.W.2d paragraph 16 (emphasis added).

[50] Northpointe describes COMPAS's scope as follows: "Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision and case management of offenders." Northpointe, Inc., "Practitioner's Guide to COMPAS Core," 1.

[51] Skeem, Monahan, and Lowenkamp, "Gender, Risk Assessment, and Sanctioning"; DeMichele et al., "The Public Safety Assessment"; Corbett-Davies and Goel, "The Measure and Mismeasure of Fairness." Note here that there is another potential issue of responsibility and stakes and of what we will call "substantive fairness" in Section 3.5. It is indeed the case that men are much more likely to reoffend and to commit violent offenses than women, though one's gender is not a factor for which one should be held responsible. Moreover, whether gender is justifiably a difference-maker in determining *sentencing* (as opposed to, say, job training, drug and alcohol counseling, or supportive intervention) will turn on a normative theory of criminal law.

[52] Angwin et al., "Machine Bias," May 23, 2016.

that was critical of ProPublica's reporting.[53] They claimed that despite its misidentifying Black defendants as high risk at a higher rate, COMPAS was unbiased. This is because the defendants within risk categories reoffended at the same rates, regardless of whether they are Black or White.

The back and forth between Northpointe and ProPublica is at the center of a dispute over how to measure fairness in algorithmic systems. Northpointe's standard of fairness is known as "calibration," which requires that outcomes (in this case reoffense) are probabilistically independent of protected attributes (especially race and ethnicity), given one's risk score (in this case high risk of reoffense and low risk of reoffense).[54] In this context, calibration requires that knowing a defendant's risk score *and race* should provide the same amount of information with respect to their chances of reoffending as *just* knowing their score. ProPublica's standard of fairness, on the other hand, is "classification parity," which requires that classification error is equal across groups, defined by protected attributes.[55] As the dispute between ProPublica and Northpointe shows, you cannot always satisfy both standards of fairness.

This might be counterintuitive. ProPublica's conclusions about COMPAS are a result of the fact that Black defendants and White defendants are arrested and rearrested at different rates, and hence Black and White defendants are counted as "re-offending" at different rates. To better understand how COMPAS can satisfy calibration but violate classification parity, it will be helpful to substitute a version of COMPAS with simplified numbers, which we will call "SIMPLE COMPAS." Note that we choose the numbers here because they loosely approximate ProPublica's analysis of COMPAS, which included larger numbers of Black defendants counted as reoffending.

SIMPLE COMPAS. SIMPLE COMPAS sorts defendants into two risk groups: high and low. Within each group, defendants reoffend at the same rates, regardless of race. In the low-risk group, defendants reoffend about 20 percent of the time. In the high, about 80 percent. The high-risk group is overwhelmingly (but not entirely) Black, and the low-risk group is overwhelmingly (but not entirely) White. Its results are summarized by the following bar chart (Figure 3.1).[56]

Now consider three questions about SIMPLE COMPAS.

First, if you randomly select a defendant, not knowing whether they are from the high- or low-risk group, would learning their race warrant suspicion that their chance of reoffending is higher (or lower) than others from the risk group they are

[53] Dieterich, Mendoza, and Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity."

[54] Corbett-Davies and Goel, "The Measure and Mismeasure of Fairness."

[55] Corbett-Davies and Goel.

[56] We borrow the idea of using this kind of chart to relay the difference between calibration and classification parity from Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness." The image itself is similar to one used in Castro, "Just Machines."

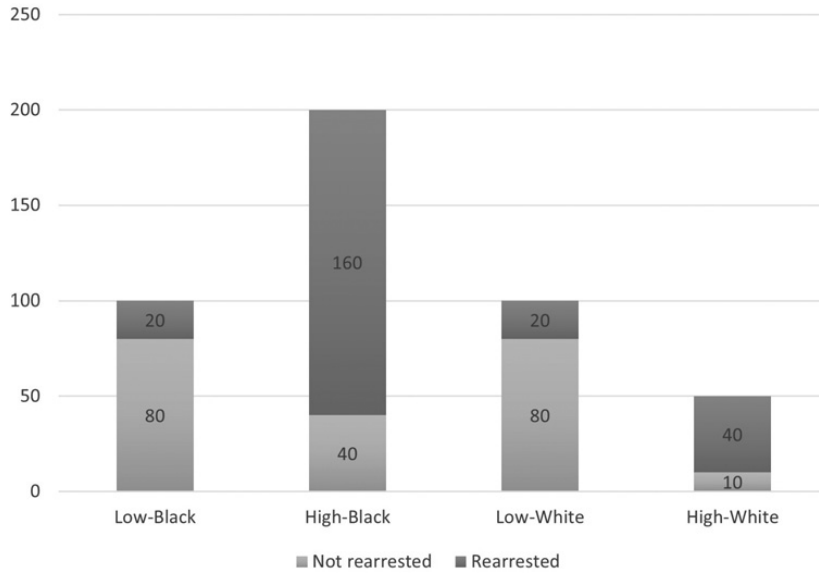

FIGURE 3.1 SIMPLE COMPAS

in? No. As we have stipulated, defendants within a risk group reoffend at similar rates regardless of race, and this is reflected in the bar chart, where the *proportion* of reoffenders to non-reoffenders in "Low – Black" is the same as "Low – White": one in five. Similarly, the proportion of reoffenders to non-reoffenders in "High – Black" is the same as "High – White": four in five. In virtue of this, SIMPLE COMPAS satisfies the calibration standard of fairness.

Second, if you randomly select a defendant, not knowing whether they are from the high- or low-risk group, would learning their race warrant increasing or decreasing your confidence that they are in the high-risk group? Yes. We have stipulated (tracking the analysis of COMPAS from ProPublica) that the high-risk group is predominantly Black and that the low-risk group is predominantly White (80 percent). If the randomly selected defendant is Black, you should increase your confidence that they are from the high-risk group, from about 55 percent (since five out nine total defendants – 250/450 – are high risk) to about 66 percent (since six out of nine Black defendants – 200/300 – are high risk). If, on the other hand, the randomly selected defendant is White, you should increase your confidence that they are from the low-risk group, from about 44 percent (since four out of nine defendants – 200/450 – are low risk) to about 66 percent (since six out of nine White defendants – 100/150 – are low risk).

Third, if you randomly select a defendant, not knowing whether they are from the high- or low-risk group, should learning their race affect your confidence that they are non-reoffending high-risk (i.e., misidentified as high-risk)? Yes. To see this, just

https://doi.org/10.1017/9781108895057.003 Published online by Cambridge University Press

look back to the previous question. If you learn the defendant is Black, you should increase your confidence that they are from the high-risk group. In virtue of this, your confidence that they are in the *non-reoffending* high-risk group should increase too, from about 11 percent (since one out of nine defendants – 50/450 – are non-reoffending high risk) to about 13 percent (since two out of fifteen Black defendants – 40/300 – are non-reoffending high risk). This may not seem like much, unless we appreciate that learning that a defendant is White should drive your confidence that they are non-reoffending high risk down to about 6.6 percent (since one out of fifteen White defendants – 10/150 – are non-reoffending high risk). This means that, in SIMPLE COMPAS, Black defendants are *twice as likely as White defendants to be misidentified as high-risk*, which violates classification parity. And this is what ProPublica found: COMPAS misidentifies Black defendants as high risk about twice as often as it does White defendants.[57] As this should make clear, violating classification parity is one way that an algorithmic system can impose an undue relative burden.

Analyzing SIMPLE COMPAS further shows why this is a matter of relative burden and why this sort of issue is distinct from issues of reliability, responsibility, and stakes. Suppose that in the world of SIMPLE COMPAS, Black and White citizens use illicit drugs at similar rates. However, Black citizens are disproportionately charged with drug crimes because they are more likely to get stopped and searched. Because of this correlation, SIMPLE COMPAS (which does not directly take race into account) is more likely to identify Black defendants as high risk.

If we assume further that the justice system that SIMPLE COMPAS is embedded in has sensible penalties,[58] then the shortcomings of SIMPLE COMPAS defy the categories of responsibility, reliability, and stakes. SIMPLE COMPAS does not hold defendants responsible for factors not under their control: The vast majority of those tagged as high risk are in fact likely to reoffend, and they are being held responsible for breaking laws that they in fact broke. Similarly, we can't complain that SIMPLE COMPAS is unreliable; it is well calibrated and predicts future arrests very reliably. Finally, we cannot complain from the perspective of stakes, because – as we have stipulated – the justice system SIMPLE COMPAS is embedded in has sensible penalties. Yet something is wrong with the use of SIMPLE COMPAS, and the problem has to do with relative burden.

To see this, compare SIMPLE COMPAS with EVEN COMPAS.

EVEN COMPAS. EVEN COMPAS sorts defendants into two risk groups: high and low. Within each group, defendants reoffend at the same rates, regardless of race. In the low-risk group, defendants reoffend about 20 percent of the time. In the high-risk group, about 80 percent . The high-risk group and low-risk group are each

[57] Angwin et al., "Machine Bias," May 23, 2016.

[58] Of course, it may well be unjust to impose criminal penalties for many kinds of drug possession and use in the first place. We will leave that issue aside for this project.

50 percent White and 50 percent Black. The sizes of the low- and high-risk groups are such that EVEN COMPAS misclassifies all defendants at the same (low) rate that SIMPLE COMPAS misclassifies Black defendants.

Suppose that EVEN COMPAS, like SIMPLE COMPAS, is embedded in a system that has sensible penalties. The only difference is that the EVEN COMPAS police force does not discriminate, and so EVEN COMPAS does not learn to identify Black defendants as high risk more often than White defendants.

Let us now make two observations. First, note that EVEN COMPAS might not be problematic. That is, given the details of the case, it does not seem like there are obvious objections to its use that we can make: It is an accurate and equitable device. It does misclassify some non-reoffending defendants as high risk, but unless we abandon pretrial risk assessment altogether or achieve clairvoyance, this is unavoidable.

Second, note that SIMPLE COMPAS is intuitively problematic even though – from defendants' points of view – EVEN COMPAS treats them worse on the whole (i.e., compared to SIMPLE COMPAS, EVEN COMPAS is worse for White defendants and it is not better for Black defendants). What could explain SIMPLE COMPAS being problematic while EVEN COMPAS is not? It cannot be an issue of reliability, responsibility, or stakes. Rather, it is the fact that SIMPLE COMPAS imposes its burdens unevenly: It is systematically worse for Black defendants. Hence, relative burden is a factor in whether people could (or could not) reasonably endorse a system that is distinct from reliability, responsibility, and stakes.

Let's return to *Loomis*. COMPAS – like the fictional SIMPLE COMPAS – does have an issue with high relative burden. The burden does not happen to be one that negatively affects Loomis. He does not have an individual complaint that he has endured a burden that is relatively higher than other people subject to it. However, it does not follow that COMPAS is a system that *any* person subject to it can reasonably endorse. Rather, because it imposes a greater relative burden to Black defendants than to White defendants, it is one that at least some defendants cannot reasonably endorse.

3.5 WHY NOT FAIRNESS?

None of the top-line criticisms of algorithmic systems that we offer in this book are that such systems are unfair or biased. This might be surprising, considering the gigantic and expanding literature on algorithmic fairness and bias. It is certainly true that decision systems are in many cases biased and (hence) unfair, and it is also true that unfairness is an extremely important issue in the justifiability of those systems. There are several, related reasons we do not primarily lean on fairness. First is that whether something is fair is often best understood as the conclusion of an argument. As we note in Section 3.3, whether people can reasonably endorse a system to which they are subject can be understood as a criterion for whether that system is fair. Likewise, the component parts of our reasonable endorsement argument can be understood as

https://doi.org/10.1017/9781108895057.003 Published online by Cambridge University Press

questions of fairness. So whether a system imposes a relative burden that is arbitrary, compound, or otherwise unjustifiable is a way in which a system can be unfair.

The second reason that we don't lead with fairness is that there is an important ambiguity in conceptions of fairness. The issue is that fairness is a concept that can in some senses be formalized, but in other senses serves as an umbrella concept for lots of more specific moral values. To see why this is important, and why we think it is fruitful not to deploy fairness as a marquee concern, we need to distinguish two broad conceptions of fairness. First is,

Formal fairness: the equal and impartial application of rules.[59]

Formal fairness contrasts with

Substantive fairness: the satisfaction of a certain subset of applicable moral reasons (such as desert, agreements, needs, and side-constraints).[60]

This distinction is straightforward. Any system of rules can be applied equally in a rote or mechanical way and, thus, can arrive at outcomes that are in some sense "fair" so long as those rules are applied without deviation. The rule that deli customers must collect a ticket upon entry and will be served in the order of the ticket numbers is a rule that can be applied in a formally fair way. However, if it is easy to steal tickets, if some tickets are not sequentially numbered, or if some people are unable to stand in line, the rule will be substantively unfair because it fails to satisfy important, applicable moral reasons that the rule does not cover. One might add all kinds of more complicated rules (people may move to the front if they need to, there will be no secondary market in low numbered tickets, people may only buy a defined, reasonable amount, etc.). Each of those additional rules may be applied in a formally fair way. Nonetheless, we cannot ensure substantive fairness by ensuring formal fairness, regardless of how exacting, equal, and impartial an application of rules is. That is because substantive fairness itself *just is* a conclusion about (a) which moral reasons are applicable and (b) whether those reasons have been proportionally satisfied. And (a) and (b) cannot be answered by an appeal to *formal* fairness without assuming the answer to the question at hand, viz., what moral reasons ought to apply.

Moreover, it might be logically impossible to simultaneously maximize different facets of substantive fairness. In other words, even if we could agree on a set of substantive moral criteria that are relevant in determining whether a given algorithmic system is fair, it may not be possible to take those substantive criteria and render a formally fair application of rules that satisfies those criteria. To understand why, consider recent work by Sam Corbett-Davies and Sharad Goel.[61]

In the literature on fairness in machine learning, there are three predominant conceptions of fairness. Corbett-Davies and Goel argue that each one is inadequate

[59] Hooker, "Fairness." See also Castro, "Just Machines."
[60] Hooker. See also Castro, "Just Machines."
[61] Corbett-Davies and Goel, "The Measure and Mismeasure of Fairness."

and that it is impossible to satisfy them all at once. The first is "anti-classification," according to which algorithms do not take into consideration protected characteristics (race, gender, or close approximations). They argue that anti-classification is an inadequate principle of fairness, on the grounds that it can harm people. For example, because women are much less likely than men to commit violent crimes, "gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions."[62] Notice that their argument implicitly incorporates a substantive moral theory about punishment, namely that justification for punishment for violent crimes depends on the likelihood of the criminal committing future offenses. Hence, the question of whether "anti-classification" is the right measure of fairness requires addressing a further question of substantive fairness.

The second main conception of fairness in machine learning Corbett-Davies and Goel describe is "classification parity," which requires that predictive performance of an algorithm "be equal across groups defined by . . . protected attributes."[63] We explained earlier how ProPublica's examination of COMPAS showed that it violates classification parity. The problem is that when distributions of risk *actually* vary across groups, achieving classification parity will "often require implicitly or explicitly misclassifying low-risk members of one group as high-risk, and high-risk members of another as low risk, potentially harming members of all groups in the process."[64]

The third conception, calibration, requires that results are "independent of protected attributes after controlling for estimated risk." One problem with calibration is the reverse of classification parity. Where there are different underlying rates across groups, calibration will conflict with classification parity (as we discussed earlier). In addition, Corbett-Davies and Goel argue that coarse measures that are well calibrated can be used in discriminatory ways (e.g., by using neighborhood as a proxy for credit-worthiness without taking into account income and credit history).[65]

The upshot of the Corbett-Davies and Goel paper is that the results of using each of the formal definitions of fairness are in some way harmful, discriminatory, or otherwise unjustifiable. But that is simply another way of saying that there are some relevant, applicable moral claims that would not be proportionally addressed in each. In other words, there may be substantive reasons that measures of formal fairness are good. However, because substantive fairness is multifaceted, no single measure of formal fairness can capture it.

Another line of literature attends to the fact that there are different conceptions of substantive fairness in the philosophical literature, each of which has different

[62] Corbett-Davies and Goel, 2.
[63] Corbett-Davies and Goel, 2.
[64] Corbett-Davies and Goel, 2.
[65] Corbett-Davies and Goel, 2–3.

implications for uses of algorithmic systems.[66] The fact that there are different conceptions of fairness and that those different conceptions prescribe different uses and constraints for algorithmic systems is largely a function of the scope of substantive fairness. That is, substantive fairness is capacious, and different conceptions of fairness, discrimination, egalitarianism, and the like are component parts of it.

Finally, a conclusion that a system is fair will often be tenuous. That is because a system that renders outcomes that are formally and substantively fair in one context may be rendered substantively unfair when the context changes.[67] Consider our discussion of COMPAS. We can imagine a risk assessment tool that is strictly used for interventions meant to prevent violence and reoffense with, for example, drug and alcohol treatment, housing, job training, and so forth. Such a system could (let's suppose) be formally fair and proportionally address relevant moral reasons. However, if that same system is deployed in a *punitive* way, then a different relevant moral reason is applicable, namely that it is substantively unfair to punish people based on facts for which they are not blameworthy. Addictions, housing insecurity, and unemployment are not conditions for which people are blameworthy. Hence, the new application of the risk algorithm would be substantively unfair, even if the original application is not.

To sum up, substantive fairness is broad and includes a range of relevant moral reasons. The purpose of this project is to examine one important component of relevant moral reasons. Thus, we don't begin with fairness.

3.6 CONCLUSION

Our task in this chapter has been to link our conception of autonomy and its value to moral principles that can serve as a framework for when using algorithmic systems is justifiable. We did so by arguing for the Reasonable Endorsement Test, according to which an action is morally permissible only if it would be allowed by principles that each person subject to it could reasonably endorse. In the context of algorithmic systems, that principle is that subjecting a person to an algorithmic decision system is morally permissible only if it would be allowed by principles that everyone could reasonably endorse. From there, we offered several factors for when algorithmic systems are such that people subject to them can reasonably endorse them. Specifically, reasonable endorsement is a function of whether systems are reliable, whether they turn on factors for which subjects are responsible, the stakes involved, and whether they impose unjustified relative burdens on persons.

[66] Binns, "Fairness in Machine Learning: Lessons from Political Philosophy."
[67] Herington, "Measuring Fairness in an Unfair World."

Notice, though, that these criteria are merely *necessary* conditions for permissibility based on respect for persons. They are not sufficient. For example, use of algorithmic systems may meet these criteria but will not be justifiable for other reasons. Indeed, one common criticism of algorithmic systems is that they are inscrutable (either because the technology is complex or because access is protected by intellectual property laws). We consider that in Chapter 4.