What If Algorithmic Fairness Is a Category Error? [1]

Arvind Narayanan

August 26, 2025

Accepted for publication. Cite as: Narayanan, A. (2026). What if algorithmic fairness is a category error? In S. Nyholm, A. Kasirzadeh, & J. Zerilli (Eds.), *Contemporary debates in the ethics of artificial intelligence* (pp. 77–96). Wiley-Blackwell. ISBN 9781394258819.

Algorithmic decision-making systems don't exist in a vacuum. To meaningfully assess their harms, we have to examine them in the organizational and political contexts in which they are deployed. Thus, in my view, the debate about whether we can make algorithms fair misses the mark — the statistical properties of algorithms tend to have at best a nebulous relationship with real-world outcomes for people. Furthermore, the lens of discrimination is inadequate to understand the plethora of harms from algorithmic decision-making.

I advocate for a more ambitious study of fairness and justice in algorithmic decision making in which we attempt to model the sociotechnical system, not just the technical subsystem. The animating question becomes: “How should we design algorithmic bureaucracies?” This will require many shifts including letting go of neat, mathematically precise fairness definitions and embracing empirical social scientific methods. But the potential payoff is enormous in terms of a greater ability to model benefits and harms and much expanded design space for reform.

A The Limits of Algorithmic Fairness

Algorithmic fairness as a movement, in my reckoning, is a little over a decade old in 2025. (As an idea, it is much older [Ochigame 2020].) It has been spearheaded by a coalition of civil rights groups, academics, and journalists. It has exercised strong pressure on policy makers, regulators, and companies. Since my interest is in whether algorithmic decision making can be fair in practice, rather than in theory, I find it helpful to start by considering the track record of this movement.[2]

[1] I developed the ideas in this essay through various talks and lectures: “Bias bias, and other biases,” a keynote at the 2023 European Workshop on Algorithmic Fairness, “Is Algorithmic Fairness Even Possible?” a talk at a Future of Privacy Forum event in 2024, and a course on “Limits to Prediction” in Spring 2024 that I co-taught with Matthew J. Salganik. I'm grateful to those audiences for feedback. I also thank Solon Barocas, Sayash Kapoor, and Aleksandra Korolova for comments on a draft of this essay.

[2] My focus in this essay is specifically on algorithmic decision-making systems, although the algorithmic fairness community has a broader scope including, for instance, generative artificial intelligence (AI) systems.

There has been an avalanche of research on how to make algorithms fair, and there has been a powerful movement to turn those ideas into reality. How have things panned out?

I will argue that this movement has been only minimally effective at preventing harms from automated decision-making systems. When we analyze why, it reveals two important limitations of the underlying ideas. First, fairness as a proxy for justice focuses attention on too narrow a set of questions. Second, it applies a depoliticized lens that gives an illusion of moral clarity in academic discussions but runs into headwinds when actually attempting to implement it. These attributes are not incidental and cannot easily be fixed. They are integral to what makes the fairness frame appealing in the first place.

A few words about what I mean by fairness as a frame. Consider ProPublica's “machine bias” investigation, famous for bringing attention to injustices in criminal risk prediction tools by showing that a prominent tool, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), had nearly twice the false positive rate for Black defendants as White defendants (Angwin et al. 2016). (A false positive is when a defendant flagged as high risk does not go on to recidivate.) The reporting, the ensuing flood of academic research, and advocacy from civil society groups all took this racial disparity as their primary focus.

But this is not the only way to question whether the system is harmful. Let us think of a counterfactual world where the investigation did not reveal a racial disparity. Would we consider the use of COMPAS in such a world to be just?

Pretrial detention based on risk is a form of precrime – incarcerating people not as a way to punish past crimes but as a way to pre-empt future crime (Zedner 2007). There is a long-standing debate about its legitimacy, separate from any questions about racial disparities. Precrime is arguably unjust because it undermines the state's commitment to respecting individuals as responsible agents and shifts the conception of justice from rehabilitation and restoration to incapacitation (among other reasons) (Duff 2013; Harcourt 2007). My point is simply that this is an important debate to be had, not a particular view on the debate.

There are many other normative objections to risk prediction systems. By being hard to understand and challenge, they may deprive defendants of procedural protections. They create disparate burdens on communities that have a larger share of incarcerated people due to spillover effects (Huq 2019). They may act as bandages that forestall deeper reforms.

In short, one might object to criminal risk prediction and pretrial detention on the basis of disparities in treatment or outcomes of defendants *relative* to each other, or on the basis that when the state uses them, it doesn't uphold what it owes to *any* defendant, or we might criticize the system as a whole, casting the net more widely than their effects on defendants. (For an exposition of the difference between the first two notions, see Hellman 2016.)

My central point is that the first frame has come to dominate. Of course, there is plenty of work on the other frames as well, but the conception of justice in terms of relative notions of discrimination has been the original, central, and galvanizing frame in the algorithmic fairness community – not just in the domain of criminal justice but in hiring, education, welfare, and every other area in which algorithmic decision making systems have proliferated. The fairness frame has become ritualized in the proliferating practice of "bias audits."

Within this frame, there are many axes of variation in definitions of fairness: disparities in treatment versus outcomes, individuals versus groups, and so on. These differences do not undercut my point.

A.1. The Appeal of Algorithmic Fairness

The algorithmic fairness community in the United States consists of civil rights groups, consumer groups, scholars in many disciplines including computer science, law, and philosophy, journalists, data-and-justice think tanks, progressive lawmakers, algorithm auditors, DEI (diversity, equity, and inclusion) consultants, and more.[3] The coalescence of such a varied set of actors may seem obvious in retrospect but is in fact remarkable and deserves an explanation.

There are two things that are not a-priori obvious. The first, as we've discussed, is that fairness would become the dominant concern regarding the harms of algorithmic decision making systems. The second is that a community would coalesce at all and remain stable over a long period of time.

When this community was my intellectual home, none of this seemed surprising to me nor, as far as I am aware, to most other members of the community. Co-authoring *Fairness and Machine Learning* was a half-decade-long exercise in disillusionment (the subtitle is *Limitations and Opportunities*) (Barocas et al. 2023b) that led to my moving on to other areas, which have led me to ask these questions now, with the benefits of time, hindsight, and distance, looking in from the outside.

The Advocacy Coalition Framework is a helpful way to understand the algorithmic fairness movement (Sabatier and Jenkins-Smith 1993). Advocacy coalitions are made up of actors who come together because of their shared beliefs (rather than primarily their material interests).

The core belief in this case is that technology must serve the ends of justice and not just profit. But there are many ways in which this high-level belief can be expressed and turned into action. Why has fairness been the dominant frame?

[3] Some of the analysis and many of the examples in this essay are US specific, since that is where I am based, but the lessons that I draw should be broadly applicable.

I think there is no single answer. The fairness frame is appealing to different groups in the coalition for different reasons, some of which I will now discuss. Forgive my broad brush – while there are of course exceptions to the patterns I highlight below, they are meant to explain the behavior of (sub)communities as a whole, which is the relevant unit of analysis, and not individuals.

Mathematical tractability: Statisticians, computer scientists, and engineers are drawn to fairness because it allows them to set aside the social and organizational context in which an algorithmic decision making system is deployed and instead formulate a clean, tractable technical problem. (This is the framing trap [Selbst et al. 2019].)

Legal hooks: Similarly, fairness is appealing to legal scholars because it generates a font of new problems under well-established bodies of law. Figuring out how antidiscrimination law in employment, credit, and many other areas should apply to automated systems is far from clear a-priori. These puzzles have led to much productive legal scholarship. In contrast, if we center other perspectives such as the need for reforming hiring processes (which I argue subsequently would be more fruitful), the role of legal scholarship is less clear. Relatedly, businesses have a strong incentive to avoid liability under anti-discrimination law.

Moral intuition: Discrimination concerns readily trigger moral intuitions in people in a way that other equally serious concerns, such as inadequate procedural protections, often don't. This has helped journalists make a powerful case to the public about the risks of algorithmic decision making and raise the salience of the frame. For example, the ProPublica investigation of COMPAS generated copious academic debate about whether unequal error rates even matter (Berk et al. 2018; Chouldechova 2017; Corbett-Davies et al. 2017; Dieterich et al. 2016; Hardt et al. 2016; Hedden 2021; Hellman 2020; Kleinberg et al. 2017; Long 2020; Mitchell et al. 2018/2021; Pleiss et al. 2017; Rudin et al. 2020;). COMPAS does satisfy other statistical fairness criteria such as calibration that may be incompatible with error rate parity. But this debate mattered little in terms of public perception. The disparities documented by ProPublica seemed to speak for themselves. (For the record, in *Fairness and Machine Learning* we expressed the view that error rate parity is indeed a meaningful metric, but that *all* fairness metrics are at best diagnostics that might alert us to the overall sociotechnical decision-making system being discriminatory; they should never be treated as constraints to be satisfied (Barocas et al. 2023).)

New attention to old problems: Abebe et al. point out that “Computing can foreground long-standing social problems in a new way” (Abebe et al. 2020, p. 257). Algorithmic discrimination has proved strategically useful to scholars of inequality, civil rights groups, and other actors, as a way to bring new attention to the plight of the vulnerable. One reason for this is that it is relatively straightforward to quantify bias in centralized, automated decision-making systems compared to decentralized and poorly legible traditional systems.

When long-standing issues of inequality became newsworthy for a new reason about a decade ago, it rapidly generated demand for investigation of biases in algorithmic systems. This gave rise to a feedback loop of increasing salience and awareness, through which the idea that algorithms can be biased went from a fringe notion to something approaching a consensus in a strikingly short period of time in the mid-to-late 2010s.

A back door for values: Questions of redistribution, especially in countries like the United States, tend to be highly partisan and perpetually mired in well-trodden terrain, with seemingly little hope for progress. But recasting these questions in technical terms takes them out of the public eye and raises hope for a more technocratic, expert-driven mode of progress. From a legal perspective, the fairness frame allows algorithmic justice to be understood as mere enforcement of existing antidiscrimination laws in the context of new decision making systems. This is much easier to advocate for compared to new policy and seemingly bypasses the need for political debate.

But this depoliticization is a mirage. Fairness scholars rightly point out that discrimination tends to operate in facially neutral ways, and that this potential is exacerbated by the use of algorithms. For example, employment and income-related criteria for access to credit, while facially neutral, can be considered discriminatory in a context where some groups have much more difficulty obtaining well-paying jobs than others. Often there is no truly neutral solution to thorny questions of allocating scarce resources, merely trade-offs between the needs of different individuals and groups.

But by the same token, “fixing” discrimination cannot be done in a scientifically objective, neutral manner. How do we know how strong of a fairness intervention to apply? It is inevitably a question of values, and requires political contestation. Solon Barocas presciently warned in 2017 that many algorithmic fairness efforts might be affirmative action in disguise and would meet with resistance when exposed as such (Barocas 2017).

Attempts at “depoliticization of politics” are common among advocacy communities, especially in debates where one side argues that its positions are inevitable consequences of “following the science” (Parkhurst 2017). Advocates may not recognize the extent to which these seemingly scientific positions are molded by smuggled-in values.

Co-optation and ethics washing: From the perspective of vendors and deployers of algorithmic decision-making tools, the fairness frame was much less threatening than calls for systemic reform. Companies decided to co-opt the movement rather than resist it, reframing it as AI ethics, or responsible AI. In this reframing, the critiques were watered down into a form that were compatible with the logics of corporate power, often conceptualized as vague principles that did not necessitate real change but could be satisfied through technical tweaks and self-regulatory procedural approaches somewhat disconnected from the underlying normative concerns (Bietti 2021; Birhane 2021; Greene et al. 2019; Jobin et al. 2019; Metcalf et al. 2019; Mittelstadt 2019; Schellmann 2024).

A.2. A Bandage for a Bandage

Consider hiring automation, which is one major area where algorithmic fairness concerns have arisen. Auditing for bias has become standard practice in this industry – in fact, one of the selling points of vendors is that clients who use such software to screen job applicants can have peace of mind from knowing that their products have passed bias audits (Raghavan et al. 2020).

Have these practices adequately addressed harms from hiring automation? Not really. The deeper problem is that many of the products in this category don't necessarily work in the first place (Raji et al. 2022). As far as I'm aware, vendors of these products have never opened up to independent audits of validity rather than bias, thus we don't have good evidence that machine-learning-based tools can predict outcomes of interest such as job performance.

In the absence of such evidence, there are reasons to be skeptical prima facie. Many of these tools are based on automated personality assessments conducted using methods such as one-way video interviews (in which the candidates answer predetermined questions and are assessed by AI) or games in which candidates are tested on how long they wait before popping a virtual balloon on the screen (among other such tasks) – tasks that have little discernible connection to any job-related skills.

On the few occasions in which independent researchers or investigative journalists have managed to test these tools, the results have not inspired confidence (Rhea et al. 2022). For example, Retorio, a video analysis tool, produced much higher scores for the *same* video when it was digitally altered to add a bookshelf in the background or glasses to the candidate's face (Harlan and Schnuck 2021). This is exactly what we should expect for a tool that was likely trained on data that did not contain any signals that would allow inferring job-relevant skills but plenty of features such as the presence of glasses or bookshelves that correlate with personality or job performance in superficial ways.

But even if these tools don't work – even if they are, as I have called them elsewhere, elaborate random number generators (Narayanan 2019) – are they actually causing harm, assuming that they don't produce disparities along race, gender, or another socially salient dimension? In my view, yes. The primary harm is that candidates are demeaned by being required to be assessed by an automated system and having their life opportunities be tied to their ability to perform for a robot and figure out what tricks would please it.

Ironically, these types of screening procedures are rarely used for hiring the psychologists and software engineers who build such tools; their use is concentrated in occupations such as retail and call center workers who are paid and valued relatively little. One consequence is that the harms from the use of these tools are concentrated among lower-income people. There are dozens of definitions of fairness, but this kind of discriminatory effect is invisible to any of those definitions because their scope tends to be limited to a single decision-making system.

Why do employers use these tools despite a lack of evidence of functionality? In the book *AI Snake Oil*, our hypothesis is that they simply don't care (Narayanan and Kapoor 2024). Traditional interview processes are not very effective either; in many jobs, there may not be good ways to identify who will perform well (beyond basic screening for qualifications). Thus, automated screening may not lose much in terms of candidate. The elaborate pretense of using cutting-edge AI might function as a cover to avoid confronting the fact that hiring processes are broken and companies don't know how to fix them. In addition to this psychological comfort that hiring automation brings to decision makers, it also enables significant cost savings and acts as a legal defense (Raghavan et al. 2020).

In the realm of automated tools used in social services, Virginia Eubanks's book *Automating Inequality* warned that the technology served as a distraction from the underlying problem, which is that society's approach to poverty is not working (Eubanks 2018). As our example of hiring automation shows, this phenomenon of tech-as-bandage arises not just in the public sector but also in the private sector.

Note that if the algorithm itself is a distraction from the real harms, which are systemic, the issue of algorithmic fairness is *two* levels removed from what needs to be fixed. It is a bandage for a bandage!

That said, it is less clear if the push for fairness in such domains is merely ineffectual at mitigating injustices or if it makes things worse. Although many scholars assert or imply that incremental improvements can forestall deeper reforms, this is hard to test empirically.

Nonetheless, the fact that vendors of hiring algorithms appeared to get significant public relations mileage out of bias audits, while avoiding any scrutiny of validity, suggests that these companies are able to co-opt the salience of discrimination concerns in strategic ways (HireVue 2021; Horowitz 2019; Wilson et al. 2021; Young et al. 2022; Younis 2019).

A.3. Three Root Causes behind the Persistence of Algorithmic Harms

Over the past decade there have been dozens of examples of algorithmic injustices in the public discourse. But it is surprisingly hard to identify many instances that can be understood and, more important, were adequately remedied through the lens of discrimination.

In the small number of cases where the problem can adequately be understood as a discriminatory algorithm, without the inherently political considerations I discuss later, companies and governments have generally been willing to work to fix or mitigate the issue. Examples that fit this mold to various degrees include Facebook's Variance Reduction System for its ad targeting (Timmaraju et al. 2023, but see Imana et al. 2025), Amazon's scrapping of a resume screener that was found to have a gender bias (Dastin 2018), and the UK police making changes to its Harm Assessment Risk Tool system due to concerns about disproportionately flagging individuals from poorer neighborhoods (Burgess 2018).

To be clear, none of this would have happened without public scrutiny and legal action (or the threat thereof) when necessary. Acting as a persistent, coordinated force for such accountability is a notable achievement of the algorithmic fairness community.

However, as noted, only in a small number of cases were (somewhat) happy endings even possible. When we look at the reasons why the discrimination frame has proved inadequate in the rest of the cases, three clusters emerge.

First, statistical disparities may merely be the symptom of deeper maladies, as discussed previously and emphasized by many scholars (Benjamin 2019). Fixing broken processes and institutions requires reform movements specific to those institutions, rather than a broad-based movement such as algorithmic fairness. Such movements exist in a few of the domains in question, such as criminal justice, and seem to have a mixed track record.

In a second cluster of cases, the root cause is the lack of procedural protections. I'm particularly concerned about explanation, appeals, contestability, and other such protections that apply to individual decisions in the deployment phase, rather than bias audits, impact assessments, and stakeholder consultation that primarily apply to the design and development phase.

The lack of procedural protections has been particularly common when automated systems have been used to accuse people of welfare fraud, often based on flawed data, without opportunity for appeal. To name just one example, in the Netherlands, the use of such a system in the 2010s led to a scandal that led to the resignation of the prime minister and his entire cabinet (”Dutch Government Resigns” 2021).

To be clear, the need for appeal, recourse, oversight, and other procedural protections is well recognized in the algorithmic fairness community (after all, the Fairness, Accountability, and Transparency conference has accountability in its name). But it has proved much harder to move the needle on requiring or incentivizing decision-makers to incorporate meaningful procedural protections than it has been to get them to fix statistical disparities.

One reason such protections tend to be resisted by developers and deployers is that they are costly, as they require human involvement, thus undercutting a major motivation for the turn to algorithmic decision making in the first place. We have observed a repeated pattern where vendors sell tools with the promise of full automation, but when errors inevitably arise, retreat to the fine print that says that a human must always make the final decision (Kapoor and Narayanan 2022).

The problem is that it is not obvious how to achieve the best trade-off between the efficiency benefits of automation and the need for procedural protections. As I discuss, this is an area where further research is needed.

The European Union AI Act does take a comprehensive approach to accountability in the context of “high-risk” AI systems. It remains to be seen whether this will be effective or will be met by compliance theater.

The third and final cluster of cases involves a set of normative and political considerations that resist resolution through scholarly debate. Here's an example. In 2016, Amazon used a data-driven system to determine the neighborhoods in which to offer free same-day delivery. According to Amazon, the decisions are made based on efficiency and cost considerations. But as discussed earlier, facially neutral procedures often result in discriminatory effects. An investigation revealed that White residents were more than twice as likely as Black residents to live in one of the qualifying neighborhoods (Ingold and Soper 2016).

While we don't know the full details, based on many similar examples, we may surmise what happened: the histories of racial inequality in the United States are encoded in residential segregation, income differences, and shopping patterns, thus being reflected in the decisions made by Amazon's algorithm. Amazon's purely cost-minimizing approach is unjust because it perpetuates those very inequalities.

What to do about it is less clear. What is an acceptable level of demographic disparity? That is, to what extent is it Amazon's responsibility to remedy the country's centuries of past injustice? And who gets to decide?

We can come up with a fairness definition that will provide almost any conceivable answer to the first question. My contention is that all of those definitions are of minimal relevance, because the proper venue for this debate is not scholarship. It is democratic politics. Not necessarily electoral politics, but the public sphere more broadly. To the extent that scholarship has a role, it is to change public opinion, not inform decision-making directly. No matter how brilliantly argued the philosophical justification for a particular fairness criterion, if it is not politically palatable to the public at large, it will result in fierce resistance.

At least in the United States, redistributive policies remain broadly unpopular among the public (Pew Research Center 2023), which has limited the potential for algorithmic fairness interventions. Unfortunately, the fairness community has generally shied away from the work of changing public opinion. I can only speculate about the reasons for this, but based on my experiences in the community, one reason that is sometimes given is that this approach to change takes too long and that marginalized communities have been discriminated against for long enough. Whatever the moral merits of this argument, I think it is strategically an unwise approach.

B How Should We Design Algorithmic Bureaucracies?

There are two ways to interpret the question of whether algorithms can be fair. The first is as an inert statement about the statistical properties of algorithmic subsystems of decision making systems. Unfortunately, such a framing has only a tenuous bearing on questions of outcomes, harms, and justice, although it makes for seemingly profound debates about the incompatibility between different fairness criteria. It is effectively a category error.

The other alternative is to interpret fairness as a property of the overall decision making system. But if we do, we cannot meaningfully answer it at the level of the algorithm. We need a model of the entire system, that is, the organization or institution that makes those decisions, the decision subjects, and the wider societal context. These systems can be modeled as “algorithmic bureaucracies” (Vogl et al. 2019).

Before I explain what I mean by this, let me clarify what I mean by bureaucracy. I use the term in a purely descriptive sense, without the negative connotation usually attached to it. And in my usage, it applies equally to public and private sector organizations. I use it to mean a highly rational, hierarchical decision-making system in which policies are developed by higher-level officials or decision-makers, and implemented by “street-level bureaucrats” (Lipsky 2010). For example, in the context of social services, these would be case workers who interview individuals and adjudicate applications based on predetermined policies.

Before we ask how the introduction of algorithms changes things, we need a bit more detail about how traditional bureaucracies function. Many people, especially technologists, have an intuitive mental model that goes something like this. First, policy makers specify the goals, values, and objectives of the system. Then they formulate a policy that, based on the evidence available, best achieves the specified objectives. This policy is then shipped to the line workers who implement it faithfully.

Under this mental model, automated decision-making is the pinnacle of increasing rationalization. It has nothing to say about the first step – the goals, values, and objectives are treated as exogenous to the system. But the remaining two steps, the thinking goes, can be automated. The best policy can be automatically determined using machine learning from past examples, as well as automatically implemented. The role for policy makers is merely to specify the objective function for machine learning, and there is no role for street-level bureaucrats.

Unfortunately, almost everything about this naive mental model is wrong. Describing its limitations is a good way to lay out a research agenda for the design of better algorithmic bureaucracies.

B.1. Values and Goals in an Algorithmic Bureaucracy

The mental model is closely related to what is called the rational-comprehensive model of policy making. In a foundational 1959 paper called the *Science of Muddling Through*, Charles Lindblom exposed the rational-comprehensive model as a myth, at least in the context of public administration (Lindblom 1959). Instead, he describes how administrators actually make policy: the values are not explicitly specified or agreed upon at the beginning, and the test of a good policy is simply that decision makers directly agree that it is good, rather than agreeing that it is the most appropriate means to an agreed objective.

I'm eliding many details here. An important point of his paper is that there is an actual process by which this happens (which he terms the method of “successive limited comparisons”), even if it may superficially appear as though there is no method at all.

Lindblom defends this approach as necessary and argues that attempting to achieve explicitness about values in a way that is divorced from the consideration of specific policies would be utterly futile:

> [T]here is no practicable way to state marginal objectives or values except in terms of particular policies. That one value is preferred to another in one decision situation does not mean that it will be preferred in another decision situation in which it can be had only at great sacrifice of another value. Attempts to rank or order values in general and abstract terms so that they do not shift from decision to decision end up by ignoring the relevant marginal preferences. ...
>
> Unable consequently to formulate the relevant values first and then choose among policies to achieve them, administrators must choose directly among alternative policies that offer different marginal combinations of values. Somewhat paradoxically, the only practicable way to disclose one's relevant marginal values even to oneself is to describe the policy one chooses to achieve them. Except roughly and vaguely, I know of no way to describe – or even to understand – what my relative evaluations are for, say, freedom and security, speed and accuracy in governmental decisions, or low taxes and better schools than to describe my preferences among specific policy choices that might be made between the alternatives in each of the pairs.

Mudding Through is brilliant and compelling, and I cannot do justice to it with a few quotes. It is a shame that it doesn't appear to be well known in the algorithmic fairness community, because it has enormous implications for how we design algorithmic bureaucracies. (One of the few efforts to make the connection is Ryan Calo's *Modeling Through* [Calo 2022]; I am grateful to Calo for introducing me to the original.)

Over 70 years later, muddling through still appears to be the norm. Rebecca Johnson and Simone Zhang published an excellent analysis of bureaucratic policies as a baseline for understanding what the turn to algorithmic decision-making might offer (Johnson and Zhang 2022). Though their focus is on describing the policies themselves, which they term "categorical prioritization," their description of the underlying process, notably the intertwining of values and policies, is consistent with bureaucrats muddling through.

Here's the problem. Algorithmic decision-making, at least as traditionally conceptualized, requires specifying goals and values up front in a stand-alone way, but in practice decision-makers lack a principled way to do so (Wang et al. 2024).

This issue is an Achilles' heel of algorithmic decision-making, and the literature on problem formulation or problem formalization considers this question (Levy et al. 2021). It makes it hard to incorporate a multiplicity of values, and instead, whichever party is most persistent and persuasive may be able to prioritize their preferred goals. Worse, specifying the objectives might be deferred to the technologists building the model, which is deeply problematic in public-sector contexts because they do not have the proper authority to make these normative decisions (Citron 2008). Worst of all is when those technologists end up making those decisions based on whatever is most convenient from a technical perspective (Passi and Barocas 2019).

For example, consider the shift to algorithmic decision-making in organ transplant allocation (Narayanan et al. 2024). As usual, there is a multiplicity of values, some utilitarian and some deontic. We may want to make decisions in a way that maximizes the number of lives saved through transplantation (or life-years added to recipients), but we may also recognize that some recipients are more deserving than others (for example, genetic diseases versus avoidable lifestyle choices leading to organ failure). These are just two of the many considerations. It turns out that algorithmic decision-making has privileged utilitarian considerations above others, and this massive shift in medical ethics appears to have happened without much public awareness or debate.

An urgent question in the design of algorithmic bureaucracy, then, is to develop processes for explicit specification of goals that

- can be programmed into automated decision-making systems (whether encoded as objective functions for machine learning or in some other manner)
- reflect a considered compromise between stakeholders and their values

- is legible to the public, allowing for oversight and minimizing the risk of smuggled-in values (Lazar 2024; Barocas et al. 2023)

This process must recognize the difficulties highlighted by Lindblom. It may not be possible to carry out value deliberation and goal specification in a truly a priori fashion, and it might have to be integrated with model building, evaluation, and explanation, so that analysts can consider how the trade-offs among values and goals are reflected in concrete policies implemented through machine-learned models. It might be hard to achieve this when the decision-making is performed entirely by a predictive model. For this reason, it will be important to find ways to effectively hybridize categorical and algorithmic prioritization (Wang et al. 2024).

Table 6.1 Comparison of how traditional and algorithmic bureaucracies handle the three levels of decision making

	Values and goals	**Policies**	**Adjudication**
Traditional bureaucracy	Intertwined with policy making; not made explicit	Categorical prioritization	Street-level bureaucrats
Algorithmic bureaucracy	Must be made explicit	Machine learning models and/or rules implemented in software	Combination of human and automated system

B.2. Opportunities for Cost–Benefit Analysis

One curious limitation of the current design of algorithmic bureaucracies is that while some parts are highly rationalized and optimized – notably the use of machine learning to optimize predictive accuracy, sometimes under statistical fairness constraints – other aspects are left up to intuition, informal deliberation, or qualitative evidence.

Thus, while I have criticized overquantification in the previous section, at the same time there is an opportunity for introducing quantified approaches, notably cost–benefit analysis, in some aspects of the design of such systems. But any application of cost–benefit analysis must pay heed to long-standing critiques such as the difficulty of valuing nonmarket goods such as freedoms, the need for considering distributional effects rather than merely total costs and benefits, and the risk of smuggled-in values.

Still, many striking findings have been obtained through quantifying and comparing previously unquantified aspects of decision-making systems.

One such aspect is the decision of where to set the risk threshold for classification. For example, in a pretrial detention application, at what threshold of predicted risk should defendants be detained? This is a hard question to answer because the benefits (public safety) and costs (to freedom) seem incommensurable. Thus, most of the literature on comparing different policies (such as different algorithms or human judges vs algorithms) skirts this question entirely.

For example, in a landmark paper comparing human decisions and machine predictions, Kleinberg et al. write about the use of algorithmic prediction: “[O]ne policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates … these gains can be achieved while simultaneously reducing racial disparities” (Kleinberg et al. 2018). They say that how this trade-off between crime and detention is made depends on judges' and society's preferences but argue that regardless of where exactly that preference lies, risk prediction would offer an improvement over the status quo. Of course, there is nothing wrong with this argument, but it does leave an important question unanswered.

But in the paper “Pretrial Detention and the Value of Liberty,” Megan Stevenson and Sandra Mayson confront the trade-off. Their method is simple: asking survey respondents to “choose between being the victim of certain crimes or being jailed for varying time periods.” What they find is that “even short periods of incarceration impose grave harms, such that a person must pose an extremely high risk of serious crime in order for detention to be justified. No existing risk assessment tool is sufficient to identify individuals who warrant detention” (Stevenson and Mayson 2022). Along similar lines, Anderson et al. use quantitative methods to compare the experiences of people in pretrial detention and those in prison post conviction and conclude that pretrial detention is “inextricably punitive” (Anderson et al. 2024). In general, findings of this nature may support deeper reform efforts and they may also be helpful for the more incremental yet highly significant questions about how to set thresholds.

Another way in which the question of thresholds is frequently skirted is to assume that there is a fixed resource that is to be allocated. For example, in the domain of social services, one may assume there is a fixed budget for welfare or assistance, and the decision to be made is simply to choose which eligible individuals should be prioritized for assistance.

But nothing is truly fixed. It costs money to collect data and build a model, and there are potentially other, harder-to-quantify costs of algorithmic decision making such as privacy. What if some of that budget were instead to be allocated to expanding access, so that more recipients benefitted, at the expense of the model being slightly worse at identifying the most needy? Under certain conditions, Juan Carlos Perdomo shows that welfare can indeed be maximized using this simpler approach (Fischer-Abaigar et al. 2025; Perdomo 2024).

Findings about the trade-off between access and accuracy or the relative value of liberty and safety can certainly be contested; different methods or assumptions may yield different results. But the general point is that finding clever ways to quantitatively compare seemingly incommensurable goals or values can open profoundly insightful lines of inquiry for further investigation.

In *Fairness and Machine Learning*, we pointed out that there is a plethora of organizational interventions that can improve fairness without touching the algorithmic decision-making system, such as expanding outreach efforts to underrepresented groups in a hiring process. The trade-offs between cost, fairness, and other desiderata introduced by such policies have not been studied with nearly the same zeal as accuracy-fairness tradeoffs. This is an opportunity for future research; it will require empirical data for applying econometric methods and/or advances in realistic theoretical modeling and simulation of various systems

In predictive modeling, it is common to assume that the benefits/costs of positive/negative classification are the same for all decision subjects. But in practice, this might be far from being true. In a college dropout risk prediction task, one student might be at risk of dropping out because they are struggling academically and another because they don't value college and plan to quit in order to pursue a venture-funded technology startup. Obviously, the former would benefit much more from academic counseling and assistance than the latter. To be sensitive to such considerations, we must formulate intervention problems rather than prediction problems, allowing us to combine the power of machine learning with methods from causal inference. Liu et al. provide an overview of this emerging science (Liu et al. 2025).

Another often-ignored consideration is spillover: decisions about some subjects could have important impacts on others, whether positive (as in the case of vaccination) or negative (as in the case of incarceration). In a setting where the supply of vaccines is scarce (such as early in the COVID-19 pandemic when vaccine production was still ramping up), a study argued that vastly more lives can be saved by allocating them to the most socially active individuals rather than to the oldest individuals (Chen et al. 2021).

Of course, this finding by itself doesn't imply that such a policy is a good idea – it raises immediate questions about moral hazard (incentivizing people to be more promiscuous) and deontic considerations (society's duty to protect the most frail). In “Against Predictive Optimization,” we argued that the actual vaccine allocation policy adopted in the United States did notably well at accommodating a multiplicity of values and goals (Wang et al. 2024). Still, it is important to ask whether spillover effects can be modeled in decision-making without compromising other desiderata.

One type of cost that I have not seen many attempts to quantify is that of human involvement in algorithmic decision-making, whether for oversight, appeals, or any other aspect. If a small increase in accuracy of a model led to a big increase in this cost because of, say, lower interpretability and explainability, the trade-off would not be worth it.

Similarly, it is also important to quantify the cost of complying with regulations, though this is already common (European Commission 2021). I must again acknowledge the dangers of this approach – cost–benefit analysis has a long history of being wielded as an antiregulatory tool because the costs of regulations are often easier to quantify than the benefits (Pasquale 2021). Still, it is important to know if the same effects can be achieved with lower compliance costs.

B.3. Adjudication in an Algorithmic Bureaucracy

Another way in which the aforementioned naive model falls short is in treating street-level bureaucrats as mere rule-following human automata and thus suitable for replacement by machine learning as long as the model performs well enough in terms of accuracy, fairness, and other relevant metrics.

There are very few decision-making systems where this is the case. One notable example is content moderation on social media. Social media companies seem to strip out the discretion of human moderators to the maximum extent possible and compel them to adhere to minutely specified policies that attempt (yet fail) to anticipate every possible circumstance. As we wrote in *AI Snake Oil*:

> Viana Ferguson, a former Facebook content moderator, recalled an incident where she encountered a picture of a White family with a Black child, captioned “a home is not a home without a pet.” She felt it was plainly racist and dehumanizing: There was no pet in the image, and there was no doubt about what the word pet was in reference to. Although dehumanizing speech is against Facebook's policies, Ferguson could not convince her manager that the post should be taken down, apparently because there was nothing in the rules that applied when the effect was achieved through the combination of the image and the caption. Moderators are “paid to follow orders, not think.” (Oremus 2020, n.p.)

This example shows why street-level bureaucrats are usually given significant discretion despite the well-known concern that they might abuse that discretion. No policy can anticipate every possible exigent circumstance. Indeed, when street-level bureaucrats are not given discretion or don't exercise that discretion, we experience them as cold, devoid of common sense, and even cruel, which is a big part of the reason why the term bureaucracy has a negative connotation.

Research has repeatedly shown that the way in which algorithmic risk scores ultimately impact decision subjects is heavily moderated by the behavior of street-level bureaucrats. For example, the introduction of risk assessment in Kentucky “benefited white defendants more than blacks.

However, this is not because the risk assessment was more racially biased than judicial discretion. Rather, it is due to regional differences in how judges responded … Judges from predominantly white rural counties liberalized their bail setting practices [by incorporating risk assessment] more than judges from more racially mixed urban areas, but *within* the same county, white and black defendants saw similar increases in release" (Stevenson 2018, footnotes omitted).

The need for discretion, common sense, and adapting to novel situations remains very much true when implementing algorithmically generated policies rather than traditional ones – if anything, even more so. When judges override algorithmic risk assessments, it is often because of important moral considerations such as the lower culpability of younger defendants that were not taken into account in the design of the algorithm (Stevenson and Doleac 2024). In fact, street-level bureaucrats will go so far as to manipulate the inputs to an algorithmic decision-making system in the interest of ensuring justice (Raso 2017).

Given this evolving understanding of the intricate interplay between human decision-makers and algorithms, a major unsolved problem is to design "street-level algorithms" (Alkhatib and Bernstein 2019). Alkhatib and Bernstein' describe the core challenge in their seminal work:

> When street–level bureaucrats encounter a novel or marginal case, they use that case to refine their understanding of the policy. When street–level algorithms encounter a novel or marginal case, they execute their pre–trained classification boundary, potentially with erroneously high confidence. For a bureaucrat, but not an algorithm, the execution of policy is itself reflexive. For an algorithm, but not for a bureaucrat, reflexivity can only occur after the system receives feedback or additional training data. The result is that street–level algorithms sometimes make nonsensical decisions, never revisiting the decision or the motivating rationale until it has prompted human review.

To the extent that the goal in deploying an automated decision-making system is to minimize humans in the loop and human oversight, an expansive conception of street-level algorithms is necessary. (This is not to say that the goal itself is always appropriate; nor do I mean to imply that eliminating rather than merely reducing the need for human oversight might be possible.)

A street-level algorithm is not merely a predictive model that spits out a score, but a unified agent that handles all aspects of interaction with the decision subject, retaining a memory of past interactions. It is able to provide, when the subject requests it, various kinds of explanations: of the overall logic behind the decision system, of any specific decisions it makes about the subject, of the subject's rights and appeal procedures, and so forth. It is able to reflexively adapt policies to novel situations or escalate to a human when it is unable to do so. It is able to handle some subset of appeals, such as when a subject states that data about her in the system is erroneous and provides supporting documentation. It is also able to handle some types of recourse, such as advising rejected loan applicants on options available for improving their prospects (Ustun et al. 2019).

To be clear, in 2025, street-level algorithms are not within close reach. It requires progress in many active research directions such as generating explanations that are both faithful and understandable. It will likely involve some combination of large language models and special-purpose models, but today's large language models have many limitations: they are susceptible to going off the rails when receiving adversarial inputs, their actual understanding of policies may lag their ability to fluently simulate such understanding, and their affective aspects raise many concerns including sycophancy and manipulation.

Note that studying and improving fairness will look completely different when considering street-level algorithms compared to traditional predictive models. It will no longer be possible to do so in an isolated way based on statistical properties of algorithms. Instead, it will be necessary to study the interaction between the system and the decision subjects. A broader tool kit will be required.

Today, many predictive tools are built and advertised without clarity on whether they are meant to replace or help human decision-makers. This is a mistake. Furthermore, given that street-level algorithms (as I have conceived of them) don't yet exist, fairness requires that predictive tools focus on decision support instead for the time being.

Here, again, the statistical properties of the underlying models are only one small factor that affect how fairness will play out in practice. At least as important is the human–computer interaction aspect of the system: How well does the system help the human understand and correct implicit bias? Does it lead to the user's skills improving or degrading over time? Do the users show algorithm aversion? (Dietvorst et al. 2015) or automation bias? (Parasuraman and Manzey 2010). How good are the explanations accompanying the model predictions? Do they give the user enough information to know when to override the recommendation? (Note that explanation for expert users is a much more tractable problem than explanations for nonexperts/decision subjects.)

There is a long way to go in terms of designing effective complementarity between human decision-makers and algorithmic decision support tools (Vaccaro et al. 2024). For example, in one study, 10% of judges outperformed algorithms predictions when performing a discretionary override, and they seem to use private information not available to the algorithmic tool in order to do so (Angelova et al. 2023). But there is much to understand about exactly how these high-skill judges operate and how tools can be redesigned to make it easier for other judges to improve their performance.

B.4. Concluding Thoughts

I have critiqued the tendency in much of the algorithmic fairness scholarship to consider questions that constitute only a narrow slice of the multitude of harms arising from algorithmic decision-making systems and to do so in a way that abstracts away many of the messy realities of these systems. This approach has enabled scholars to pursue rigor, focus on aspects of the problem that can be tackled by their respective disciplines, and allowed intellectual progress to be decoupled from political debate. But these advantages are also drawbacks.

This narrowness has long been recognized, and the “second wave of algorithmic accountability” took a more structural approach (Pasquale 2019). But this wave has had less engagement with the technical specifics of algorithmic systems compared to the first wave, limiting it to somewhat blunt (albeit vitally important) questions such as which systems should be built at all.

In between these contrasting approaches lies a vast space that constitutes more meaningful reform than tweaking the weights of models and doesn't shirk the inherently political nature of the task at hand, while still embracing quantification, technical specificity, and emerging AI capabilities. Perhaps it is time for a third wave.

References

Abebe, R., Barocas, S., Kleinberg, J. et al. (2020). Roles for computing in social change. In: *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260. New York: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372871

Alkhatib, A. and Bernstein, M. S. (2019). Street-level algorithms: a theory at the gaps between policy and decisions. In: *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, 1–13. New York: Association for Computing M et al. achinery. https://doi.org/10.1145/3290605.3300760

Anderson, C. N., Cochran, J. C., and Montes, A. N. (2024). How punitive is pretrial? measuring the relative pains of pretrial incarceration. Punishment & Society 26 (5): 790–812. https://doi.org/10.1177/14624745231218702

Angelova, V., Dobbie, W. S., and Yang, C. (2023). Algorithmic Recommendations and Human Discretion (NBER Working Paper No. 31747). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w31747

Angwin, J., Larson, J., Mattu, S. et al. (2016). Machine bias: there's software used across the country to predict future criminals. And it's biased against blacks. ProPublica, 23 May.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 22 August 2025).

Barocas, S. (2017). What is the problem to which fair machine learning is the solution? AI Now Conference, 10 July. https://www.youtube.com/watch?v=S_AkPi6-r3Y (accessed 22 August 2025).

Barocas, S., Hardt, M., and Narayanan, A. (2023a). When is automated decision making legitimate? (chap. 2). In: Fairness and Machine Learning: Limitations and Opportunities, 44–75. Cambridge, MA: MIT Press. https://fairmlbook.org/classification.html (accessed 22 August 2025).

Barocas, S., Hardt, M., and Narayanan, A. (2023b). Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA: MIT Press. https://fairmlbook.org (accessed 22 August 2025).

Barocas, S., Hardt, M., and Narayanan, A. (2023c). Relative notions of fairness (chap. 4). In: Fairness and Machine Learning: Limitations and Opportunities, 76–103. Cambridge, MA: MIT Press. https://fairmlbook.org/relative.html (accessed 22 August 2025).

Benjamin, R. (2019). Race after Technology: Abolitionist Tools for the New Jim Code. Cambridge: Polity.

Berk, R., Heidari, H., Jabbari, S. et al. (2018). Fairness in criminal justice risk assessments: the state of the art. Sociological Methods & Research 50 (1): 3–44.

Bietti, E. (2021). From ethics washing to ethics bashing: a moral philosophy view on tech ethics. Journal of Social Computing (Spec. issue “Technology Ethics in Action”) 2 (3). https://ieeexplore.ieee.org/document/9684746 (accessed 22 August 2025).

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. Patterns 2 (2): 100205. https://doi.org/10.1016/j.patter.2021.100205

Burgess, M. (2018). UK police are using AI to inform custodial decisions—But it could be discriminating against the poor. Wired, 1 March. https://www.wired.com/story/police-ai-uk-durham-hart-checkpoint-algorithm-edit (accessed 22 August 2025).

Calo, R. (2022). Modeling through. Duke Law Journal, 71 (6): 1391–1423.

Chen, J., Hoopes, S., Marathe, A. et al. (2021). Prioritizing allocation of COVID-19 vaccines based on social contacts. medRxiv [Preprint]. Feb 16:2021.02.04.21251012. https://doi.org/10.1101/2021.02.04.21251012.

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5 (2): 153–163.

Citron, D. K. (2008). Technological due process. Washington University Law Review 85: 1249–1314.

Corbett-Davies, S., Pierson, E., Feller, A. et al. (2017). Algorithmic decision making and the cost of fairness. In: KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 797–806. New York: Association for Computing Machinery. https://doi.org/10.1145/3097983.3098095

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 10 October. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG (accessed 22 August 2025).

Dieterich, W., Mendoza, C., and Brennan, T. (2016). COMPAS risk scales: demonstrating accuracy, equity, and predictive parity. Northpointe (Equivant) white paper. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf (accessed 22 August 2025).

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144 (1): 114–126. https://doi.org/10.1037/xge0000033

Duff, A. (2013). Pre-trial detention and the presumption of innocence. In: Prevention and the Limits of the Criminal Law (ed. A. Ashworth, L. Zedner, and P. Tomlin), 115–132. Oxford: Oxford University Press. https://academic.oup.com/book/33065/chapter/281701798 (accessed 22 August 2025).

Dutch government resigns over child benefits scandal. (2021). The Guardian, 15 January. https://www.theguardian.com/world/2021/jan/15/dutch-government-resigns-over-child-benefits-scandal (accessed 22 August 2025).

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press. https://us.macmillan.com/books/9781250074317/automatinginequality (accessed 22 August 2025).

European Commission. (2021). Commission staff working document: Impact assessment accompanying the proposal for a regulation of the European Parliament and of the Council laying down 22armonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (SWD(2021) 84 final). EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021SC0084 (accessed 22 August 2025).

Fischer-Abaigar, U., Kern, C., and Perdomo, J. C. (2025). The value of prediction in identifying the worst-off. "." Poster presented at the Forty-Second International Conference on Machine Learning, Vancouver, BC, 13–19 July. https://icml.cc/virtual/2025/poster/46605 (accessed 22 August 2025).

Greene, D., Hoffmann, A. L., and Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical AI/ML. In: Proceedings of the 52nd Hawaii International Conference on System Sciences, 2122–2131. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1261&context=hicss-52 (accessed 22 August 2025).

Harcourt, B. E. (2007). Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age. Chicago: University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/A/bo4101022.html (accessed 22 August 2025).

Hardt, M., Price, E., and Srebro, N. (2016).Equality of Opportunity in Supervised Learning. In: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems (ed. D. D. Lee, U. von Luxberg, R. Garnett et al.), 3323–3331. Red Hook, NY: Curran Associates. https://dl.acm.org/doi/10.5555/3157382.3157469 (accessed 22 August 2025).

Harlan, E. and Schnuck, O. (2021). Objective or biased: How AI evaluates job applicants. Bavarian Broadcasting, 16 February. https://interaktiv.br.de/ki-bewerbung/en (accessed 22 August 2025).

Hedden, B. (2021). *On statistical criteria of algorithmic fairness*. Philosophy & Public Affairs 49 (2): 209–231. https://doi.org/10.1111/papa.12189.

Hellman, D. (2016). Two concepts of discrimination. Virginia Law Review, 102 (4): 895–952. https://virginialawreview.org/articles/two-concepts-discrimination (accessed 22 August 2025).

Hellman, D. (2020). *Measuring algorithmic fairness*. Virginia Law Review 106 (4): 811–866.

HireVue. (2021). HireVue leads the industry with commitment to transparent and ethical use of AI in hiring. Press release, 12 January. https://www.hirevue.com/press-release/hirevue-leads-the-industry-with-commitment-to-transparent-and-ethical-use-of-ai-in-hiring (accessed 22 August 2025).

Horowitz, J. M. (2019). Americans see advantages and challenges in country's growing racial and ethnic diversity. Washington, DC: Pew Researcj Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2019/05/Views-of-diversity_FINAL_05.08.19.pdf (accessed 22 August 2025).

Huq, A. Z. (2019). *Racial equity in algorithmic criminal justice*. Duke Law Journal 68 (6): 1043–1134. https://scholarship.law.duke.edu/dlj/vol68/iss6/1 (accessed 22 August 2025).

Imana, B., Shen, Z., Heidemann, J., & Korolova, A. (2025). External evaluation of discrimination mitigation efforts in Meta’s ad delivery. Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT), Athens, Greece, June 23–26. https://doi.org/10.1145/3715275.3732170

Ingold, D. and Soper, S. (2016). Amazon doesn't consider the race of its customers. Should it? Bloomberg, 21 April. https://www.bloomberg.com/graphics/2016-amazon-same-day (accessed 22 August 2025).

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence 1: 389–399. https://www.nature.com/articles/s42256-019-0088-2 (accessed 22 August 2025).

Johnson, R. A. and Zhang, S. (2022). What is the bureaucratic counterfactual? Categorical versus algorithmic prioritization in U.S. social policy. In: FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1671–1682. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533223

Kapoor, S. and Narayanan, A. (2022). The bait and switch behind AI risk prediction tools. AI Snake Oil, 22 November. https://www.aisnakeoil.com/p/the-bait-and-switch-behind-ai-risk. (accessed 22 August 2025).

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In: ITCS 2017: 8th Innovations in Theoretical Computer Science Conference (ed. C. H. Papadimitrou), 43.1–43.23. Wadern: Dagstuhl Publishing. https://cs.emis.de/LIPIcs/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43_.pdf (accessed 22 August 2025).. ITCS

Kleinberg, J., Lakkaraju, H., Leskovec, J. et al. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics 133 (1): 237–293. https://doi.org/10.1093/qje/qjx032

Lazar, S. (2024). Legitimacy, Authority, and Democratic Duties of Explanation. In D. Sobel & S. Wall (Eds.), Oxford Studies in Political Philosophy, Vol. 10, pp. 28–56. Oxford: Oxford University Press.

Levy, K., Chasalow, K. E., and Riley, S. (2021). Algorithms and decision-making in the public sector. Annual Review of Law and Social Science 17: 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Lindblom, C. E. (1959). The science of “muddling through.” Public Administration Review 19 (2): 79–88.

Lipsky, M. (2010). Street-Level Bureaucracy: Dilemmas of the Individual in Public Services (Expanded ed.). New York: Russell Sage Foundation.

Liu, L. T., Raji, I. D., Zhou, A. et al. (2025). Bridging prediction and intervention problems in social systems. [Preprint]. arXiv:2507.05216. https://doi.org/10.48550/arXiv.2507.05216

Long, R. (2020). Against false positive-rate equality as a measure of fairness. [Preprint]. arXiv: 2007.02890. https://doi.org/10.48550/arXiv.2007.02890

Metcalf, J., Moss, E., and boyd, d. (2019). Owning ethics: corporate logics, Silicon Valley, and the institutionalization of ethics. Social Research 86 (2): 449–476.

Mitchell, S., Potash, E., Barocas, S. et al. (2018/2021). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. Annual Review of Statistics and Its Application 8: 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902 (arXiv preprint, 2018, https://doi.org/10.48550/arXiv.1811.07867)

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI Nature Machine Intelligence 1: 501–507. https://arxiv.org/abs/1906.06668 arXiv

Narayanan, A. (2019). How to recognize AI snake oil [Talk slides]. MIT Program in Science, Technology, and Society. https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf (accessed 22 August 2025).

Narayanan, A. and Kapoor, S. (2024). AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton, NJ: Princeton University Press.

Narayanan, A., Wang, A., Kapoor, S. et al. (2024). Does the UK's liver transplant matching algorithm systematically exclude younger patients? AI Snake Oil, 11 November. https://www.aisnakeoil.com/p/does-the-uks-liver-transplant-matching (accessed 22 August 2025).

Ochigame, R. (2020). The long history of algorithmic fairness.. Phenomenal World, 30 January. https://www.phenomenalworld.org/analysis/long-history-algorithmic-fairness (accessed 22 August 2025).

Oremus, W. (2020). Facebook's contracted moderators say they're paid to follow orders, not think. OneZero, 28 October. https://onezero.medium.com/facebooks-contracted-moderators-say-they-re-paid-to-follow-orders-not-think-40331991c6ee

Parasuraman, R. and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. Human Factors 52 (3): 381–410. https://doi.org/10.1177/0018720810376055

Parkhurst, J. O. (2017). The Politics of Evidence: From Evidence-based Policy to the Good Governance of Evidence. London: Routledge.

Pasquale, F. (2019). The second wave of algorithmic accountability. LPE Project, 25 November. https://lpeproject.org/blog/the-second-wave-of-algorithmic-accountability (accessed 22 August 2025).

Pasquale, F. (2021). Cost-benefit analysis at a crossroads: The future of quantitative policy evaluation. LPE Project, 27 September. https://lpeproject.org/blog/cost-benefit-analysis-at-a-crossroads-the-future-of-quantitative-policy-evaluation (accessed 22 August 2025).

Passi, S. and Barocas, S. (2019). Problem formulation and fairness. In: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 39–48. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287567

Perdomo, J. C. (2024). The relative value of prediction in algorithmic decision making. In: ICML ’24: Proceedings of the 41st International Conference on Machine Learning (ed. R. Salakhutdinov, Z. Kolter, K. Heller et al.), 40439–40460. JMLR.org. https://dl.acm.org/doi/10.5555/3692070.3693711 (accessed 22 August 2025).

Pew Research Center. (2023). More Americans disapprove than approve of colleges considering race/ethnicity in admissions decisions. https://www.pewresearch.org/politics/2023/06/08/more-americans-disapprove-than-approve-of-colleges-considering-race-ethnicity-in-admissions-decisions (accessed 22 August 2025).

Pleiss, G., Raghavan, M., Wu, F. et al. (2017).On fairness and calibration. In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems (ed. U. von Luxberg, I. Guyon, S. Bengio et al.), 5684–5693. Red Hook, NY: Curran Associates. https://dl.acm.org/doi/10.5555/3295222.3295319 (accessed 22 August 2025).

Raghavan, M., Barocas, S., Kleinberg, J. et al. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In: *FAT* '20: Proceedings of the 2020 Conference on*

Fairness, Accountability, and Transparency, 469–481. New York: Association for Computing Machinery.https://doi.org/10.1145/3351095.3372828

Raji, I. D., Kumar, I. E., Horowitz, A. et al. (2022). The fallacy of AI functionality. In: FAccT '22: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533158

Raso, J. (2017). Displacement as regulation: new regulatory technologies and front-line decision-making in Ontario works. Canadian Journal of Law and Society 32 (1): 75–95.
Rhea, A. K., Markey, K., D'Arinzo, L., Schellmann, H., Sloane, M., Squires, P., & Stoyanovich, J. (2022). Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 572–587). New York: ACM. https://doi.org/10.1145/3514094.3534189

Rudin, C., Wang, C., and Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction*. Harvard Data Science Review 2* (1).

Sabatier, P. A. and Jenkins-Smith, H. C. (1993). Policy Change and Learning: An Advocacy Coalition Approach. Boulder, CO: Westview Press.

Schellmann, H. (2024). The Algorithm: How AI Decides Who Gets Hired, Monitored, Promoted, and Fired—And Why We Need to Fight Back Now. New York: Grand Central Publishing. https://www.grandcentralpublishing.com/titles/hilke-schellmann/the-algorithm/9780306827365 (accessed 22 August 2025).

Selbst, A. D., Boyd, D., Friedler, S. A. et al. (2019). Fairness and abstraction in sociotechnical systems. In: FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68). New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Stevenson, M. (2018). Assessing risk assessment in action. Minnesota Law Review*,* 103: 303–376.

Stevenson, M. T. and Doleac, J. L. (2024). Algorithmic risk assessment in the hands of humans*.* American Economic Journal: Economic Policy 16 (4): 382–414.

Stevenson, M. T. and Mayson, S. G. (2022). Pretrial detention and the value of liberty. Virginia Law Review 108: 709–780.

Timmaraju, A. S., Mashayekhi, S., Chen, M. et al. (2023). Towards fairness in personalized ads using impression variance constraints. In: KDD '23: Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 4937–4947. New York: Association for Computing Machinery. https://doi.org/10.1145/3580305.3599916

Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In: FAccT '19: *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, 10–19. New York: Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3287560.3287566 (accessed 22 August 2025).

Vaccaro, M., Almaatouq, A., and Malone, T. W. (2024). When combinations of humans and AI are useful: a systematic review and meta-analysis. Nature Human Behaviour 8: 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vogl, T. M., Seidelin, C., Ganesh, B. et al. (2019). Algorithmic bureaucracy. In: dg.o 2019: Proceedings of the 20th Annual International Conference on Digital Government Research (ed. Y.-C. Chen, F. Salem, and A. Zuiderwijk), 148–153. New York: Association for Computing Machinery. https://doi.org/10.1145/3325112.3325240

Wang, A., Kapoor, S., Barocas, S. et al. (2024). Against predictive optimization: on the legitimacy of decision-making algorithms that optimize predictive accuracy ACM Journal on Responsible Computing, 1(1), Article 9. https://doi.org/10.1145/3636509

Wilson, C., Ghosh, A., Jiang, S. et al. (2021). Building and auditing fair algorithms: a case study in candidate screening. In: FAccT '21: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445928

Young, M., Katell, M. A., and Krafft, P. M. (2022). Confronting power and corporate capture at the FAccT conference. In: FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1375–1386. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533194

Younis, M. (2019). As redress for slavery, Americans oppose cash reparations. Gallup, 29 July. https://news.gallup.com/poll/261722/redress-slavery-americans-oppose-cash-reparations.aspx? (accessed 22 August 2025).

Zedner, L. (2007). Pre-crime and post-criminology? Theoretical Criminology 11 (2): 261–281. https://doi.org/10.1177/1362480607075851