Canadian Journal of Philosophy (2022), 52: 1, 26–43
doi:10.1017/can.2022.3

ARTICLE

The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems

Kathleen Creel*[,1] and Deborah Hellman[2]

[1]McCoy Family Center for Ethics in Society, Institute for Human-Centered Artificial Intelligence, Department of Philosophy, Stanford University, Stanford, CA, USA, and [2]University of Virginia School of Law, Charlottesville, VA, USA
*Corresponding author. Email: kcreel@stanford.edu

Abstract
This article examines the complaint that arbitrary algorithmic decisions wrong those whom they affect. It makes three contributions. First, it provides an analysis of what *arbitrariness* means in this context. Second, it argues that arbitrariness is not of moral concern except when special circumstances apply. However, when the same algorithm or different algorithms based on the same data are used in multiple contexts, a person may be arbitrarily excluded from a broad range of opportunities. The third contribution is to explain why this systemic exclusion is of moral concern and to offer a solution to address it.

Keywords: Automated decision-making; machine learning; fairness; opportunity; arbitrariness; hiring; lending

1. Introduction

A hiring manager faces a towering stack of resumes from which she must choose a short list of candidates to interview. A lender flips rapidly through files, evaluating the loan-worthiness of applicants based on decades of credit reports and income statements. When time is short but information ample, humans struggle to decide quickly and fairly.

Private companies and state agencies alike often turn to automated decision-making systems to aid or replace their overburdened human experts. These systems offer the promise of ease and consistency.[1] Their creators describe them as unbiased substitutes for faulty human judgment, although whether they do help to minimize human bias is controversial.[2]

Automated decision-making systems implemented in public life are typically highly standardized. Few private companies or state agencies have the technical capacities to build their own automated decision-making systems. Many companies and state agencies rush to the same private providers: over one-third of the Fortune 100 companies use the same automated candidate screener, Hirevue (Green 2021).

Given this standardization, one algorithmic decision-making system can replace or influence thousands of unique human deciders. Each of the humans so replaced had their own set of

[1]Raghavan et al. (2020) discuss whether and to what extent algorithmic hiring assessments are validated, consistent, or unbiased.

[2]For discussion of algorithmic bias, see among others Ferguson (2017), Chouldechova (2017), Safiya (2018), and Eubanks (2018).

decision-making criteria.[3] Some criteria were morally and epistemically good. Some criteria were perniciously biased. And some criteria were merely arbitrary. An individual hiring manager, for example, might choose to interview only candidates wearing purple shoelaces or hire only those who enjoy puns. Is such *arbitrariness* of moral concern? This is the first question we address, focusing on algorithmic decision-making in domains that do not provide specific criteria to guide decision-making. We argue, perhaps counterintuitively, that isolated arbitrary decisions are not of moral concern except when *other* rights make nonarbitrariness relevant.[4]

The second question we address is whether arbitrariness becomes problematic at scale. The single automated decision-making system that replaces a thousand human decision makers has its own set of good, biased, and arbitrary criteria. Since no decision-making system is perfect, some qualified people will be misclassified by any system. If the same algorithms produced by the same companies are uniformly applied across wide swathes of a single domain—hiring or lending, for example, thereby homogenizing decision outcomes—a person could be irrationally excluded from a significant number of important opportunities. This, we argue, becomes a problem of moral concern.

Importantly, this harm is likely to persist even when the automated decision-making systems are "fair" on standard metrics of fairness.[5] The algorithm might mandate that a score is equally predictive for all groups protected by antidiscrimination law; that a certain percentage of applicants from an underrepresented group are chosen; that false positive or false negative rates for these different groups of people are equalized; or that attributes such as race or gender are not relied upon, explicitly or implicitly, in decision-making.[6] Any one of these forms of fairness could be maintained while the algorithm nevertheless persistently mischaracterizes the same individuals.

We will argue that arbitrariness is sometimes morally problematic, but not for the reasons that other scholars posit. In our view, the problem with arbitrary algorithms is not fundamentally their arbitrariness but the systematicity of their arbitrariness. Automated decision-making systems that make uniform judgments across broad swathes of a sector, which we might call *algorithmic leviathans*[7] for short, limit people's opportunities in ways that are morally troubling. This moral problem will plague both arbitrary and nonarbitrary algorithms, but arbitrary algorithms have no instrumental justification. We thus recommend solutions aimed at disrupting the systematicity of algorithmic leviathans rather than the arbitrariness of their decision-making.

At this point, the reader may be wondering about the relationship between arbitrariness and systematicity. We argue that the arbitrariness of algorithms is not of moral concern and that the systemic exclusion they often yield is. What connects these claims? The answer lies in the fact that the problematic nature of systemic exclusion is often attributed to its arbitrariness, a confusion we hope to dispel. In addition, while we argue that individuals have no right that employers and lenders, to take two common examples, make decisions about hiring and lending that are nonarbitrary, a nonarbitrary hiring and lending decision has at least some reason in its favor.

[3]We focus here on replacement but note that in some cases an algorithmic decision-making system that merely plays an advisory role can lead to outcomes with more disparate impact than either human or algorithmic decision-making alone (see Albright 2019).

[4]For example, other rights make nonarbitrariness relevant in the context of criminal-justice context. Reliance interests, as when an applicant has attained a degree under the understanding that an employer or job category demands it, might also ground a claim to nonarbitrary decision-making.

[5]There has been an explosion of literature on fairness in machine learning. For the purposes of this paper, we will remain neutral as to a metric of fairness. For an overview and discussion see Gilpin et al. (2018) and Mehrabi et al. (2019). For further discussion, see Hardt, Price, and Srebro (2016), Sánchez-Monedero, Dencik, and Edwards (2020), and Hellman (2020).

[6]Whether one could eliminate the ways in which an algorithm implicitly relies on protected traits is controversial; see Hu and Kohler-Hausmann (2020).

[7]The term is introduced in König (2020). We use the term as it is defined here, influenced by the interpretation in Gandy Jr. (2020, 2021).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

This weak justification becomes relevant when arbitrariness occurs at scale because the harm produced by systemic exclusions requires justification.

The arguments in this article proceed as follows. First, we present an argument against the view that an isolated arbitrary decision morally wrongs the individual whom it misclassifies. Second, we argue for the view that when arbitrariness becomes systemic, as it is likely to do when particular algorithmic tools dominate markets and thereby homogenize decision outcomes, new and important moral problems arise. Finally, we propose technically informed solutions that can lessen the impact of algorithms at scale and so mitigate or avoid the wrong we identify.

2. Do arbitrary decisions wrong individuals?

We begin with the claim that arbitrary decisions wrong individual persons. In order to assess this claim, we need to get a clear picture of what makes a decision *arbitrary.* As we show in this section, there are many possible ways to understand a claim that a decision is arbitrary. After isolating the interpretation of this assertion that seems to best cohere with familiar complaints and which seems apt for the context of algorithmic decision-making, we argue that the complaint that an algorithmic decision is arbitrary lacks moral force.

2.a Arbitrary decision-making as unpredictable, unconstrained, or unreasonable

When a person asserts that a decision is arbitrary, what might she mean? To start, here are three different understandings of arbitrariness. First, an arbitrary decision might be one that is unpredictable. Second, an arbitrary decision might be one that is unconstrained by ex ante rules or standards. Third, an arbitrary decision could be a decision that is unreasonable or nonsensical. Let's consider each in turn. In doing so, we focus on the contexts of employment and lending, and by extension, on the decisions of employers and lenders. We focus on these two specific contexts in order to make the discussion concrete and easy to follow, but we contend that our arguments about whether arbitrary decisions wrong individuals can generalize beyond these contexts.

Perhaps arbitrary is a synonym for unpredictable and the complaint of arbitrary decision-making is the claim that a decision is unfairly unpredictable. If so, it rests on the assertion that a prospective employee or borrower has a right to a predictable hiring or lending process. People surely have an interest in the ability to plan that is made difficult when employers and lenders act unpredictably. But do they have a right to this security? Sometimes they might. Imagine an employer has listed the qualifications she seeks in a job announcement. In such a case, the prospective employee might have a right to expect that the employer make decisions on the basis of the announced criteria. But what grounds this right? Reliance might. Perhaps the fact that the prospective employee puts time and energy into applying for this job rather than another because she thinks she is better qualified for it given the announced criteria is sufficient to create an entitlement that the employer use those criteria in her selection process.

Ultimately, we are unsure about whether reliance (at least in most cases) is sufficient to limit the ability of an employer to change her focus.[8] But even if it were, the claim to predictable decision-making by the employer rests not on a general right to nonarbitariness but instead on the specific claim generated by the reliance (or perhaps by a promise) implied in the announcement of job qualifications.

Perhaps arbitrary is a synonym for unconstrained and the claim of arbitrary decision-making is a complaint that a decision is unfairly untethered to any rules or standards with which it must

[8]Imagine that the employer interviews three candidates, each of whom are strong with regard to the qualifications announced in advance. Upon interviewing these candidates, the employer now sees that her understanding of the qualifications required for the job was mistaken. She decides to hire someone with different qualities altogether. In this scenario, it does not appear that the employer has wronged any of three original candidates despite the fact that they relied on the announced job description.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

comply. This sense of arbitrariness is likely inapt in the context of algorithmic decision-making. An algorithm is, after all, rule based. The *Oxford English Dictionary* defines an algorithm as "a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer." Thus, algorithmic tools are unlikely to be criticized as arbitrary because of their lack of constraint.

But maybe this conclusion is too quick. The assertion that a decision is arbitrary in the sense of "rule-bound" may be understood to require that the rule that governs the decision be one that is understandable or reproducible.[9] Consider the demand for an understandable decision first. An algorithmic decision might be rule-governed but complex in a way that makes it difficult or impossible for the person affected to understand what the basis for that decision is. If so, one can imagine the prospective employee or borrower asserting that the decision is arbitrary.

It is easy to see why a job or loan seeker would want to be able to understand the rule that governs whether or not she is offered the job or loan. But does she have such an entitlement? Suppose that an employer was to make clear that he will hire whomever he wants. There may well be a method to his madness such that his decisions are rule-bound, but it may also be the case that neither he nor the applicants can discern what this is. It is not clear that such a procedure wrongs the applicant. A critic of our view might point out that laws sometimes require an explanation[10] and that this legal requirement supports the idea that a decision must be based on explainable reasons. While this is surely a possible view, the right to an explanation can also be supported in different ways. First, it may help to guard against improper reasons affecting decisions (race, sex, etc.). Second, it may allow prospective borrowers to make decisions in their lives that make them more able to secure loans in the future. Third, it may allow borrowers to contest inaccurate data. For example, if the lender explains that Jane Doe was denied a loan because she had previously filed for bankruptcy and she has not, the explanation requirement allows Jane to correct the data on which the algorithm relies.

In each of these cases, the function of the explanation is not to ensure that the real rule which determines who gets hired or gets a loan is in fact a comprehensible rule, as this conception of arbitrariness would require. Rather, the demand for an explainable rule works prophylactically to screen out prohibited factors, aids applicants by pointing them to things they can improve, and helps ensure that the real rule is applied to accurate facts. Much more could be said about the normative function of explanations in these contexts.[11] Our aim here is to suggest that the reasons that support explainability take us quite far from notions of arbitrariness.

The suggestion that arbitrariness as lack of constraint requires instead that a decision be reproducible is likely to be satisfied in the context of algorithmic decision-making. To the extent that an algorithm governs the decision, it will produce the same result when run on the same inputs. If the algorithm contains a degree of randomness within it (as we propose later), it still is reproducible at a higher level of abstraction.

Perhaps the worry is not that the algorithm will not yield the same result for a particular input—a job or loan applicant, for example. Rather, the worry might be that for two relevantly similar applicants, it will not yield the same result. We address this concern later when we consider whether arbitrariness should be understood as a failure to *treat like cases alike.*

Third, and most plausibly, we can understand the claim to nonarbitrary decisions as a claim that decisions that affect us should be reasonable or sensible. If so, *arbitrary* is a synonym for *unreasonable* or *irrational.* This seemingly plausible idea is open to several variations that we explore in the next section.

[9]We are grateful to Seth Lazar for pressing this point.

[10]For example, in the United States lenders are required to provide explanations for loan denials. See the Equal Credit Opportunity Act, 15 U.S.C. section 1691(d).

[11]Indeed, one of the papers in this volume takes on that project. See Vredenburgh (Forthcoming) and also Vredenburgh (2021).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

2.b Unreasonable as not good for the public, not good for the person affected, or not good for the decision-maker

A complaint that a decision is arbitrary could be understood as a complaint that the decision is unreasonable or irrational. If so, we must determine when a decision is reasonable or rational. There are at least three possible answers to that question. First, perhaps the decision must be based on a reason must be public oriented. *Yes, you didn't get the loan. But that is because if people with your qualifications received loans, there would be so many defaults on loans that the housing market would collapse and that would hurt everybody* (or something along these lines). Second, perhaps the decision must be based on a reason that relates to the affected individual and her interests. *If we gave you this loan, you would likely default and that would be bad for you. You would ruin your credit and set back your ability to attain the financial stability you seek.* Third, perhaps the decision need only be based on a reason that relates to the aims or interests of the employer or lender herself. *I didn't give you the loan because I am unlikely to make a profit by giving you a loan, and I want to make money in this business.* Of these three possibilities, we can see that the first and second understanding of what nonarbitrariness requires are more demanding[12] than the third.

This demandingness is a problem. In order to argue that a lender who denies a loan must do so for a public-oriented reason or because doing so serves the borrower's interests, we must suppose that the lender has a duty to society or to would-be borrowers to act for their interests. While it is not impossible that such a duty exists, more would need to be said to support it. More importantly, *if* such a duty exists—say, a duty for a lender to make lending decisions that are good for society overall—a failure to comply would be wrongful but its wrongfulness would relate to this failure to act in the interests of society rather than to the arbitrariness of its decision. Another way to put this point would be to say that when we understand arbitrariness as acting without good reason and interpret "good reasons" as reasons that serve society, the violation of this duty makes arbitrariness drop out of the equation altogether. The same point would hold with regard to the requirement that a lender act in the interests of the would-be borrower.

We are thus left with the idea that a decision is arbitrary when it does not serve the prospective lender or employer's own interests and is sufficiently rational and nonarbitrary when it does. In other words, lenders act nonarbitrarily when they deny would-be borrower's loans in order to maximize their own profits. In our view, this understanding of arbitrariness coheres best with the ordinary meaning of the complaint of arbitrary decision-making made by prospective employees and borrowers.

A lender or employer might act arbitrarily in this sense in two ways. First, the lender might act for no reason at all. Second, the lender might act for a reason that does not serve her own interests. We next consider each of these possibilities.

2.c Arbitrary decision-making as decision-making for no reason at all or as lacking instrumental rationality

We begin with the idea that a lender, for example, acts arbitrarily when she acts for no reason at all. This idea might mean that the lender does not formulate a policy or plan and instead spontaneously decides whether to grant or deny loans when a borrower presents herself. This sense of 'arbitrariness' returns us to the earlier contrast between the arbitrary and the rulelike or regularized. And, as we noted earlier, algorithmic decision-making is not arbitrary in this manner. But an algorithm could rely on a randomization procedure. If so, the agent decides whom to select for the treatment at issue—like whom to grant a loan—by lottery. Interestingly, lotteries are often seen as

[12] *Demandingness* is used here as it is used in the discussion of consequentialist ethics: as a measure of how much sacrifice the moral theory requires or what cost the agent would incur in complying with it. For discussion of this term see Mulgan (2001).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

paradigmatically fair procedures. Still, a person might object when an agent decides by lottery whether to give him an opportunity, benefit, or burden.

The idea behind such a complaint might be as follows: *if you are going to deny me something important, there should at least be some reason for doing so, even if the reason is simply that denying me the opportunity is good for you. But if you deny me the loan for no reason, as a random procedure like a lottery would do, this denial fails to take my interests seriously.* Is the complainant correct about this?

The first point to note is that the adoption of a randomization procedure itself might be supported by a reason. A lender or employer might think that the fairest way to allocate loans and jobs is by lottery. In this case, the lender or employer has a reason for its action—a particular view of fairness. This reason might not be one that the applicant shares and is likely not a reason that serves the lender's or employer's own interest, but it is a reason that we can understand and appreciate. As a result, this complaint is best reconstructed as the assertion that the lender or employer acts wrongly when it doesn't adopt procedures for selecting applicants that serve its own interests, which we consider below.

We now turn to the claim that a lender or employer acts arbitrarily in a morally troubling way when it adopts a selection procedure that is not well-suited to its own aims. For example, if the lender's goal is to make as much money as possible, the lender should want to identify those borrowers who will enable it best to achieve this goal. If the algorithmic tool it adopts undermines or frustrates this aim, would-be borrowers are subject to arbitrary decision-making on this understanding of that claim.

Before we evaluate this claim, we should point out that there is some further ambiguity about the complaint here as well. The borrower could be arguing that the algorithmic tool will generally do a good job identifying those borrowers most likely to repay loans and take sensible risks, etc., but erred in her particular case. Or, alternatively, the borrower could be asserting that the algorithmic tool does not serve its purported goal even in the aggregate.

2.d Lacking means-ends rationality in the aggregate or in the individual case, in fact-relative or evidence-relative terms

First, consider the claim that the algorithmic tool misclassifies an individual. This is a claim that the rule-based decision is either over-inclusive, under-inclusive, or both. While the individual under consideration is a good risk, she is nonetheless denied a loan. But is the tool arbitrary according to the definition at issue? It is not. Remember, an arbitrary decision is one that does not serve the interests of the lender or employer. While it seems that this decision is arbitrary according to this definition because this borrower would be a good risk, that answer is too quick. The lender wants to make money and thus wants to adopt a screening tool that makes her the most money. In order to claim that the lender acts arbitrarily, the borrower must show that there is another selection method that the lender could adopt that would do a better job of meeting her goals than the one she has adopted. Because any tool is likely to be imperfect, merely showing it has erred in an individual case does not show it is arbitrary on this definition. Instead, the claim must rest on how the tool does over the run of cases, which takes us to the second formulation of the claim.[13]

Here the complainant asserts that the actor uses a sorting method that is ill-suited to her aims when assessed in the aggregate. Arbitrariness, under this conception, is a claim that the selection procedure lacks means/ends or instrumental rationality. The lender's action is unlikely to achieve her own goals, and this failure of instrumental rationality has consequences for the borrower. Given what is at stake for the borrower, the lender should at least be using a tool that is well-suited to her

[13]See Schauer (2003).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

own purposes, so the argument goes. In our view, we have now neared the heart of the complaint of arbitrary algorithmic decision-making.

Yet one more layer of ambiguity remains. Should this complaint be understood in fact-relative or evidence-relative form? Suppose there is good evidence supporting the algorithmic tool's reliability in achieving the lender's aim. Is that enough to make its use nonarbitrary? If so, then the duty of nonarbitrariness should be understood in evidence-relative form. Suppose that despite the evidence supporting the tool, it does not do a good job (or a good enough job) of achieving the lender's aim. If the obligation is evidence-relative, this would not matter. But, if the obligation is fact-relative, then the lender's use of the algorithm to determine to whom to offer loans would be arbitrary and objectionable for this reason.

First consider the more demanding, fact-relative version. If the lender adopts a tool that good evidence suggests will enable her to achieve her aim, then it makes sense for her to use it. Reasonable, rational action that is thereby nonarbitrary, according to the definition we are currently examining, is action conducive to the attainment of the agent's own aims. While the tool will not in fact serve the agent's interests, an agent generally serves her own interests by adopting policies that evidence suggests will further them. In other words, the fact-relative version of the obligation seems problematic because it directs an actor to do what evidence suggests she should not. For this reason, the evidence-relative formulation is the more plausible understanding of the claim. We consider it next.

On this interpretation, the would-be borrower asserts that the lender adopts an algorithmic tool that good evidence suggests would not achieve the lender's own goal. For this reason, denying the borrower a loan is arbitrary. We have reached the heart of the claim, but in doing so reveal what an odd claim it is. Essentially, the borrower is saying to the lender: *you have a duty to those affected by your actions to do what seems reasonably likely to be good for you.* Why would one think there is such a duty? When I am pursuing my own aims, why should I have a duty to pursue them more efficiently rather than less efficiently? When others are affected, I surely should take their interests into account and consider whether their interests should outweigh my own. But the fact that I should take the interests of others into account does not give rise to an obligation for me to serve my own interests well. There is thus no general obligation for actors to adopt procedures that are non-arbitrary.

2.e Acting for bad reasons

We argue that there is no general duty to act for nonarbitrary reasons. But we recognize and affirm that there are bad reasons on which an employer or lender might act. Were an employer to refuse to hire a qualified applicant because of her minority race, for example, this decision wrongs the applicant. It wrongs the applicant not because the decision is arbitrary, however. Indeed, using race or sex as a proxy for other traits that an employer or lender seeks is sometimes, unfortunately, instrumentally rational in the sense that race, sex, and other legally protected traits are often positively correlated with these other traits.[14] That said, prohibitions on irrational decision-making by employers, lenders, and others, or demands that these actions comply with instrumental rationality could be useful in helping to avoid or detect race or sex-based decisions. In US constitutional law, for example, it is often said that so-called "rationality review" helps to "smoke out" illegitimate governmental purposes.[15] Whether a prohibition on arbitrariness is equally useful

[14]The term *statistical discrimination* is used to refer to the rational use of race, sex, and other legally prohibited traits as proxies for other traits. The fact that the rational use of race and sex as proxies for other traits is legally prohibited provides some support for the view that instrumental rationality is unrelated to moral permissibility. See generally, Deborah Hellman (2008, chap. 5).

[15]See e.g., City of Richmond v. J. A. Croson Co., 488 U.S. 469, 493 (1989) ("the purpose of strict scrutiny is to 'smoke out' illegitimate uses of race by assuring that the legislative body is pursuing a goal important enough to warrant use of a highly

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

in the context of algorithmic decisions is unclear and we express no view about it. Either way, the relevant point for our purposes is that the irrationality of the action is not problematic in itself, but only, perhaps, because it works to screen out or smoke out decisions that are bad for other reasons.

2.f Treating like cases alike

Before we consider whether arbitrariness at scale presents different moral problems than arbitrariness in the individual case, we examine a challenge to the argument just presented. While individuals may have no moral claim that selectors adopt procedures that are well-suited to the selector's own aims, perhaps people have a related claim that the selection procedure at least *treats like cases alike.* The *treat likes alike* (TLA) principle has intuitive appeal.[16] Suppose we interpret TLA to say that cases that are alike with respect to the purpose of the selection rule should be treated alike. We should begin by noting that TLA does not require that the selection tool be rational given the goals that its designers aim to achieve. For example, if the algorithmic tool mischaracterizes *A* (who should be selected given the tool's purpose but is not), then the TLA principle directs that the tool should also mischaracterize *B* (who is like *A* with respect to the purpose of the rule). If an algorithmic selection tool does not treat likes alike, then perhaps the tool is arbitrary in a problematic way that we have not yet considered.

While the claim that a failure to treat like cases alike sounds like a new and different conception of arbitrariness, this charge actually restates one of the conceptions of arbitrariness we considered earlier. There are two ways in which we might understand the TLA principle. First, it might require the consistent application of the selection rule at issue. If so, TLA demands that the procedure be rulelike. An employer or lender cannot deploy the selection algorithm in some cases and ignore it in others. Or, if she does so, then we should describe the selection procedure at a higher level of abstraction as including both the algorithm and the human decision maker who determines whether to follow or ignore the result suggested by the algorithm. If nonarbitrariness requires rulelike consistency, then this demand is unlikely to be the problem when algorithms determine treatments, as we noted earlier.

Alternatively, the charge of failure to comply with the TLA may be aimed not at the application of the rule, but instead at its content. Here the complaint is that the rule itself, even when applied consistently, fails to treat like cases alike. But, as Peter Westen's canonical criticism of this principle makes plain, people (or other "cases") are not like or unalike others inherently (1982). Rather, they are like or unalike with respect to some purpose. A tall man and a short woman who are equally skilled at widget making are *likes* with respect to selecting the best widget makers and *unalike* with respect to choosing the best players for the men's basketball team. Suppose the short woman (Jane) is selected but the tall man (Joe) is passed over by the algorithm tasked with selecting the best widget makers. In such a case, the algorithm has failed to treat cases that are alike, with respect to the purpose of rule, as alike. As a result, the algorithm is less good than it might be at its task. If a better algorithm could identify all the people who are similarly good widget makers, this would be a better algorithm and the company looking to identify workers would better achieve its goals by adopting it. The failure to comply with the TLA principle thus asserts the familiar complaint—which we rejected earlier—that the algorithmic tool does not serve the ends of the person or institution deploying it as well as it could. Whenever this occurs, it is likely to lead to dissimilar treatment of people who are alike with respect to the tool's aims. And, as we argued above in section 2.d, such

suspect tool. The test also ensures that the means chosen 'fit' this compelling goal so closely that there is little or no possibility that the motive for the classification was illegitimate racial prejudice or stereotype.")

[16]There is a longstanding debate among legal scholars about whether the treat likes alike principle is empty. See e.g., Westen (1982). Westen ignited a vigorous exchange with this article. For some representative examples of the replies, see Chemerinsky (1983), Greenawalt (1983), Waldron and Westen (1991), and Peters (1996).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

dissimilar treatment of similar cases may be frustrating to applicants for jobs, loans, admission, etc., it is not, in our view, of moral concern.

In this section, we have canvased different ways of understanding the charge of arbitrariness, especially as lodged against algorithmic decision-making. After discussing and rejecting various possibilities, we concluded that a charge of arbitrariness is best understood as the claim that the employer, lender, or other actor has adopted a screening tool that fails to achieve its own ends as well as it might. As such, this charge has no obvious moral force. If an employer or lender has adopted an inefficient means to achieve its own ends, that is unfortunate but, without more, does not wrong others. The *more* that can make a difference includes cases where employers and lenders have made commitments to applicants to decide according to certain criteria or invited reliance on announced factors or processes.[17] In such cases, the employer or lender may wrong applicants if the selection procedure is not as promised or announced. But the reason the employer or lender wrongs applicants in such cases is not because the selection process is arbitrary but instead because they have broken their promise or invited detrimental reliance. That said, we recognize that an insistence of rational, nonarbitrary decision-making by employers and lenders can, at times, help screen out impermissible use of prohibited traits like race and sex. But, again, this argument does not establish that arbitrariness itself is of moral concern.

3. Arbitrariness at scale

In the last section, we concluded that the use by an employer or lender of an arbitrary algorithm does not wrong applicants. In such a case, the algorithm indicates that a particular individual should not be hired or offered a loan despite the fact that she is qualified. In addition, a better algorithm could have been adopted which would distinguish between qualified and unqualified applicants in a manner that better allows the employer or lender to achieve its own goals. If this algorithm is used not only by one particular employer or lender but by many, does this change the moral calculus? And if so, why?

3.a Why arbitrariness at scale arises

Arbitrariness at scale is not a hypothetical concern. It stems from one of the essential challenges of machine learning: the difficulty of ensuring that the model learns neither too little nor too much. The model that learns too much "overfits": it memorizes the training data rather than learning generalizable patterns, or it learns patterns based on the noise in the training data that fail to generalize beyond the initial context. The model that learns too little "underfits": it misses generalizable patterns present in its training data.

A model that overfits by learning accidental patterns in the data set is learning arbitrary features. Many algorithmic decision-making systems are based on opaque machine learning algorithms such as deep learning (Creel 2020; Zednik 2019). Deep learning is known to suffer from "adversarial examples," or cases in which very small perturbations of the input data can lead to failures in prediction or wildly different classifications. Ilyas et al. (2019) have argued that "adversarial examples can be directly attributed to the presence of features derived from patterns in the data distribution that are highly predictive, yet brittle and incomprehensible to humans." If this is true, then some multidimensional feature correlated with the input data "purple shoelace wearer + scratches right ear often + square forehead" may be highly predictive relative to the original training data. The arbitrary feature that is being uniformly enforced by the classifier might be one of this kind.

[17]Similarly, special contexts like criminal justice, in which the interests at stake give rise to duties of accuracy, provide distinct reasons that arbitrariness is morally problematic.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

If so, the arbitrary feature is a product of overfitting. The classifier has found features that are responsive to patterns present in the data used to train the algorithm, but not to the patterns that track the real-life phenomena the modelers intended to represent. The model's ability to successfully classify, as measured by the original developers of the algorithm during the training process, might turn out to be fully dependent on overfitting and thus on accidental features in the data. A classic example is an image classifier that heavily relies on the presence of snow to distinguish wolves from dogs, perhaps because shy wolves are often photographed with telephoto lenses across barren wintry expanses.[18] The classifier's reliance on snow reflects a real property of the data set—wolves are photographed more often in the snow—but not a property of wolves themselves. Were this classifier to be tested on photographs of wolves in a very different environment, such as rescued wolves photographed indoors while being treated at a veterinary clinic, its success at identifying wolves would plummet. More importantly for our purposes, the classifier would divide photographs into "contains wolf" and "no wolf" categories based on the presence or absence of snow, an arbitrary feature.

A more recent and real-world example is that of the automated diagnostic system that learned to diagnose pneumonia based on lung scans. Its apparent success did not generalize beyond the hospital in which it was trained because, as it turned out, all it "recognized" was the difference between lung scans produced in the hospital itself and those produced using the portable scanner in the ambulance, which stamped its images with the word PORTABLE (Zech et al. 2018). As people with pneumonia were far more likely to be transported to the hospital by ambulance and scanned while in transit, so too their scans were far more likely to be recognizably stamped.

Automated decision-making systems that overfit by learning an accidental pattern in the data set (i.e., that are arbitrary) will produce systematic disadvantage to some number of people. This is because the classifier has identified a consistent set of features that it can systematically enforce to exclude the same people, but the consistent set of features does not generalize beyond the training data. For example, perhaps every one of the small number of people in the initial training data set who attended Williams College happen to be rated as terrible employees or to have fully defaulted on loans. Job seekers and loan applicants who attended Williams might then be systematically denied despite their other qualifications. Williams attendance is thus an arbitrary reason for job application denial. It is arbitrary because reliance on this feature does not serve the employer's own interests.

Likewise, a feature in an algorithmic decision-making system might be a compound of a feature genuinely related to job performance *and* an arbitrary data artifact. Gabriel Goh calls these types of features "contaminated" (2019). In such a case, a job seeker or loan applicant might argue that although something about their data as provided to the model reliably triggers a hidden feature in the automated classifier that causes their job file to be thrown away, it is unfair because the feature is "contaminated." It too is based on overfitting and thus on an arbitrary artifact of the way the model originally learned from data and not on any feature of the person themselves. Does this mean that the algorithm treats this individual applicant unfairly? We argued in section 2 that it does not.

However, if the same algorithm is used by many employers, the effect will be not only to screen out this applicant from this particular job, which also would occur if one employer had idiosyncratic preferences or theories about what makes a good employee, but to screen the prospective employee out from employment with many other employers as well. This systemic effect matters because of the scale of the exclusion.

The conceptually simplest solution would be for the state to break up the algorithmic monopoly by making it illegal for one company or one company's algorithms to dominate an entire sector of

[18]See Ribeiro, Singh, and Guestrin (2016, 8–9). This classifier was intentionally trained to be an example of a "bad" classifier, but it represents a common trend.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

hiring or lending. We encourage this and expect that it would have a host of salutary outcomes beyond the discouragement of algorithmic leviathans.

However, avoiding monopolistic power may not be sufficient to avoid uniformity of outcome. If the machine learning algorithms train on the same or similar data, ensuring a diversity of companies and algorithms will not be enough to avoid the creation of an algorithmic leviathan. Since uniformity in outcome has been observed in the case of "Imagenet," we use this example to illustrate the point that training different algorithms on the same data can lead to the same result.

The "Imagenet" database is a collection of 14 million images, each hand labeled with the primary contents of the image, such as "dog" or "strawberry." Its maintainers host the widely respected "ImageNet Large Scale Visual Recognition Challenge," a competition to identify images from a curated subset of one thousand nonoverlapping categories. It is unfortunate but not surprising that machine learning models trained on ImageNet share similar biases. More surprising is that many of the competition-winning machine learning models have similar *artifacts*.

These artifacts stem from features of the Imagenet database rather than from the task of image recognition itself. Like any finite collection of images, Imagenet has certain peculiarities. Although it is large, it is a nonrepresentative subset of possible visual experiences of the world. Gains in performance sometimes come from capitalizing on features peculiar to the database rather than generalizable features of the image categories themselves. For example, Hendrycks et al. (2020) point out the existence of "natural adversarial examples," namely real, unaltered photographs that are misclassified with high confidence. The photos reveal that Imagenet classifiers rely heavily on cues from the background of photographs and especially on "textures." An image of a squirrel leaning on a wet stone and an image of a dragonfly perched on a cross-hatched lawn chair are both misclassified with high confidence as a sea lion and a manhole cover, respectively, because the organisms are in unusual locations with heavily textured backgrounds.

The ImageNet problem occurs in many forms of data beyond images. Indeed, we should expect it to occur *more often* in other domains than it does in image recognition. Image recognition data sets are large because despite the expense of hand labeling them, images themselves are relatively cheap to produce or to source on the internet. The ImageNet team in particular has produced a vast training corpus by paying workers on MechanicalTurk to label millions of images. The ImageNet team can use the labor of workers on MechanicalTurk because images of dragonflies and squirrels can be labeled by almost anyone, unlike images of tumors, for example.

In other domains, gathering large amounts of high-quality data is expensive, even before the expense of labeling the data. This shapes what kinds of data exist in these domains. In some domains, only a few public data sets will be available. If only a few data sets are easily available, even firms that are rivals are likely to train their machine learning models on those data sets.

For example, deep learning algorithms used for clinical and medical purposes disproportionately use data from only three states, drawing "cohorts from California, Massachusetts, and New York, with little to no representation from the remaining forty-seven states" (Kaushal, Altman, and Langlotz 2020). This is likely both because the rich hospitals in those states are the hospitals most capable of running studies using deep learning and of producing the kind and scale of data appropriate for deep learning. Kaushal et al. rightly stress that since these models have been trained on a small and unrepresentative subset of states, they are unlikely to generalize successfully to populations in the other forty-seven states. Such failures of generalization are important and often commented upon. However, a less frequently discussed factor highlighted by this example is that concentration and similarity of data leads to *standardization*. Medical diagnostics in all state will now rely on a small number of algorithms that were themselves trained on data representing a relatively small and uniform cohort.

Publicly available data sets also prompt standardization of algorithms for the same reason. Cheap, widely available data sets will cause standardization because they will be used for training or pretraining. The supply-side availability of readily accessible databases may lead companies to choose them rather than gathering their own data, thus pushing them towards standardization.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

There may be further bottlenecks depending on the domain. Trainers of automated decision-making systems for lending or hiring may face legal restrictions on what kind of data may legally be used, which will further homogenize the data-source options.

Automated decision-making systems with coordinated pressures on their data sets are likely to be highly standardized, even if the automated decision-making systems are trained by different groups. This tendency will lead not only to failures of generalization but to *systematic* failures of generalization. Algorithms trained on the same or similar data sets are likely to identify the same people, or rather the same application files, as worthy or unworthy of hiring or lending. For these reasons, we suspect that the outcomes will be coordinated at scale even when the same algorithm is not applied at scale.

3.b Why systemic exclusion matters

If arbitrary decision-making is not morally troubling at the individual level, does anything change when that arbitrariness goes to scale? Rather than one employer with a screening tool that does a poor job of identifying good employees, as the employer herself might define them, instead we have a hiring tool used by many employers that irrationally excludes some group of potential employees. Rather than one bank employing a poor method to determine who should be offered a loan, many banks relying on the same flawed data sets about potential borrowers irrationally exclude some borrowers from most prime loans. Does the scale at which the irrationality now operates make a moral difference?

One potential worry is that the people affected by such systemic arbitrariness might not be just some random group of people. Instead, they might be members of groups protected by antidiscrimination law. In other words, what appeared to be simply irrational in the individual case may turn out to be disparate impact discrimination at scale. Policies and practices that produce a disparate negative impact on vulnerable groups (like racial minorities) are morally problematic, and the presence or absence of arbitrariness is not needed to demonstrate the problem with the system. In such cases, arbitrariness is only indirectly relevant. Disparate impact can sometimes be legally justified by business necessity or some reason in this vein. When a selection process is arbitrary—meaning not well-suited to achieve the employer's or lender's own aims—it surely will be unable to meet this justificatory burden.[19]

The group of people affected by an irrational screening tool need not overlap with a vulnerable or protected group, however. In such cases, what worries us is that the systematicity of these flawed tools means that a person affected in one context will likely be affected in multiple contexts. Because the tool is used in multiple settings, such as an automated tool for determining creditworthiness, the exclusion will apply across multiple actors—lenders, in this example—such that a person negatively affected will be unable to find respite with another lender.

This exclusion from multiple opportunities is in itself of moral concern. When an individual is excluded from important opportunities—employment, loans, education, etc.—this has a significant impact on her life in ways that many different moral theories are likely to find objectionable. The problem could be that she is unable to develop an adequate range of capacities (Sen 1980, 1993; Nussbaum 2011), has insufficient opportunity for genuine freedom or autonomy (Raz 1988), is among the least well-off in ways that require remediation (Rawls 1971), or is consistently excluded from opportunities in a way that establishes a social hierarchy of esteem or domination (Anderson 1999). Exclusion from a broad swath of opportunity in an important sector of life is likely to be morally problematic.

[19]Arbitrariness would matter indirectly in the same way in subgroup or intersectional cases in which the people affected are not a legally protected group but a subset of two or more such groups—Asian American women, for example.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

We will not argue for this assertion here because doing so would require its own article- or book-length treatment.[20] However, we do note that this argument makes two features of the systemic effect relevant. First, the degree of exclusion. If the algorithmic tool is not in fact an algorithmic leviathan and is used by one or two employers only, then the exclusion is less severe. Conversely, if different tools are based on the same data sets, then even with different screening tools, we may still have an algorithmic leviathan. Second, the importance of the good, service, or domain from which one is excluded also matters. We focused on employment and lending, which are important domains of opportunity. When the algorithmic tool affects a less important area of life, the moral concern is less grave.[21]

Hiring and lending are similar to one another in that they are realms of opportunity. The offer of a job or a loan expands an individual's opportunities, giving them more choices and access to life paths than they had before. For that reason, when a single or a few companies dominate these areas we worry about the effects on people's lives, and the ways that institutions with significant social power may limit people's opportunities and produce harms that are cumulative. These same worries are present, however, when multiple companies and or lending institutions use the same (or few) algorithmic tools to screen employees or prospective borrowers. An algorithmic leviathan that successfully captured the whole hiring or lending market would establish a monopoly of opportunity.

It is important to note, however, that the monopolistic harm we call attention to—namely, being denied a significant number of opportunities—is present whether the algorithmic tool is rational or arbitrary. The systematicity gives rise to the sorts of harms that limit opportunities to the degree that require justification. Arbitrariness matters because when the screening tool is arbitrary, it lacks such justification. However, when it is rational, the reason for it may still not be sufficient when the exclusionary harms are great. After all, what makes it nonarbitrary is simply that it helps the company achieve its own aims. When the harms of the tool become truly monopolistic, even this justification may not suffice.

To recap, arbitrariness at scale is of moral concern when it produces systemic and unjustified exclusion from important opportunities. Arbitrariness matters only indirectly. Just as a screening tool that produces a disparate impact on a vulnerable group requires a justification, so too a screening tool that excludes individuals from an important domain of opportunity also requires justification. When the tool is arbitrary, it lacks this justification. The practical upshot of this analysis is that a remedy should be aimed at disrupting systematicity rather than arbitrariness. We turn to these solutions in the next section.

4. Technical solutions

How should the problem of systemic arbitrariness be addressed? In theory, we could address the arbitrariness by finding all individuals who have been misclassified and correct the error of their classification, or we could find all arbitrary features and remove them. However, this would be procedurally difficult. If it were possible to algorithmically identify the individuals who had been misclassified, they would not have been misclassified in the first place. We could address the algorithmic error by adding a "human in the loop" to check the results, perhaps by evaluating a randomly selected subset of the classifications using a different method. This is equivalent to adding a second and separate classification method—namely, classification by human judgment—and applying it to a randomly selected subset of the results. In fact, adding a second classification

[20]See e.g., Fishkin (2014).

[21]If an algorithmic leviathan were able to touch multiple domains of opportunity, however, the moral concern would be more grave. For example, Fourcade and Healy discuss credit-scoring algorithms as producing "classification situations" that limit opportunities across many domains of life such as hiring, lending, insurance, and more, resulting in "a cumulative pattern of advantage and disadvantage" (2013, 559).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

method can be accomplished without a human in the loop, and this strategy will be part of our positive proposal. But recall, the heart of the problem of systemic arbitrariness lies in its systematicity rather than its arbitrariness. With that in mind, we offer solutions aimed at disrupting systematicity rather than at minimizing arbitrariness.

Identifying the misclassified *groups* will be equally difficult. As we mentioned earlier, although we focus primarily on the moral dimension of standardizing a choice based on purely arbitrary features, in many cases seemingly arbitrary features will in fact correlate with axes of existing discrimination. As Barocas and Levy note, "decisions rendered on the basis of characteristics that lack social salience may still result in disparities along socially salient lines. In many cases, newly identified groups might map closely to traditional social groups because many of the relevant points of similarity will be correlated with protected characteristics" (2020). Although such mapping may exist, bias against individuals who fall into more than one group that has historically been discriminated against may be more difficult to discover in opaque automated decision-making systems. For example, one proposal for a way to reduce bias in contemporary algorithmic decision-making systems is to "audit" the systems by running them through a series of tests and check for plausibility, robustness, and fairness (Raji and Buolamwini 2019). Although intersectional identities can be modeled (Bright, Malinsky, and Thompson 2016), disadvantage to intersectional groups may nevertheless be more difficult to identify using a method such as algorithmic auditing, especially if the group has few members in the data set or if one of the axes is not a historical axis of discrimination (Carbado 2013). Kearns et al. have formalized this problem and demonstrated that auditing for statistical fairness for subgroups is computationally hard in the worst case but can be reasonably approximated (2018, 4). However, the authors set a threshold on the size of the subgroup for their approximation, allowing them to ignore the misclassification of a subgroup if it is small relative to the size of the whole (2). By contrast, we are interested in the problem of persistently misclassifying even one person, our "algorithmic Job."[22]

It might be difficult for an auditor or even an affected person to prove that a complex nexus of intersecting identities caused this algorithmic decision-making system to uniformly misclassify her. However, on our framework, the issue can be redressed even if the auditing tools or explanatory resources are not sufficiently nuanced to identify the pathway by which they arose.

Machine learning methods are typically designed to produce a single, optimal model, given a problem formulation and data. However, many problems have the property of *multiplicity*: they can be solved equally well with more than one equally optimal (or near-optimal) model.[23] In choosing the single best model of the ones available, the learning system may reject many other possible models with equally good or only slightly inferior results. This choice is nonarbitrary in the sense of instrumental rationality: the learning system chooses the best model (or a best model) according to its metric of success, thereby performing as well as it can by its own lights.[24] However, "predictive multiplicity" results show us that in some cases there can be many equally optimal models that have the same statistical performance but deliver different token predictions. In these cases, there is no self-interested reason to choose one of the optimal models over any of the others.

[22]Kearns et al. find this problem uncompelling, arguing that "we cannot insist on any notion of statistical fairness for *every* subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. This simply corresponds to 'overfitting' a fairness constraint" (2018, 2). We agree that it is not unfair to misclassify an ex-post defined subgroup once, which is why we argue that the problem is systematic exclusion.

[23]See e.g., Marx, Calmon, and Ustun (2020); they cite Breiman's discussion of the Rashomon effect as a precursor to their formalization of predictive multiplicity and impossibility results in the fairness literature as fellow travelers. See Breiman (2001), Kleinberg, Mullainathan, and Raghavan (2016), Chouldechova (2017), and Corbett-Davies et al. (2017).

[24]Citron and Pasquale see the fact of token classification differences between models as itself a sign of arbitrariness. As they note, "29 percent of consumers had credit scores that differed by at least 50 points between the three credit bureaus. Barring some undisclosed, divergent aims of the bureaus, these variations suggest a substantial proportion of arbitrary assessments" (2014, 12). In addition, choosing only one of these variations to dominate the market would lack a nonarbitrary justification.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

The following hypothetical illustrates the sort of case in which this will arise. Imagine a pool of applicants for a job consisting of 1,000 people, of whom ten are the most qualified applicants and the rest are unqualified. Call the most qualified applicants Alice, Bob, Carmelita, DeAndre, Erika, etc. The most successful classifier might recommend the first nine most qualified applicants, Alice through Indira, but reject the tenth, Juan, leaving it with a 99.99 percent unweighted success rate. An equally successful classifier might accept Bob through Juan but reject Alice, leaving it with equally good 99.99 percent unweighted success. A third classifier, only slightly less successful than the other two, might reject Bob and Carmelita but accept the other eight, leaving it with a 99.98 percent success rate. Thus, the differences between the most successful classifiers might be differences in success on individuals rather than differences in optimality, as in this case where classifiers have similar performance but different false negatives.

One approach to the joint problems of multiplicity and systematicity is to intentionally introduce randomness to the classifier. Grgić-Hlača et al. (2017) have proposed creating a diverse set of classifiers and then randomly selecting one from the set for each decision, thereby creating "random classifier ensembles." Recall the case presented earlier of multiple classifiers with similar performance at identifying the best candidates for the job. Rather than arbitrarily choosing exactly one of the two best models, thereby denying Alice or Juan all employment opportunities at companies who use the model, a random classifier ensemble incorporates many such models that fall above a threshold of performance. In any token decision, a classifier is chosen randomly from among the group. Thus Alice, Bob, Carmelita, and Juan will all have the opportunity to be short-listed for job opportunities despite each being excluded by at least one of the models. Another advantage of the random classifier ensemble is that each candidate selected is chosen for a reason, namely that she was selected by a particular classifier as the best (or a best) candidate. The choice of one candidate over others is justifiable from the perspective of a model in a way that it might not be were randomness to be introduced into an individual model.

Random classifier ensembles represent one possible way to reduce the hegemonic control of a single algorithmic decider. Increasing randomness within a single algorithm is an improvement over a single algorithmic leviathan. Randomness reduces standardization and thereby the systemic exclusion from opportunity. And classifier ensembles, in particular, address multiplicity: the fact that the optimal model choice is often only minimally better than the next best alternative, or no better at all. They allow competing classifiers to exist within the same model. Furthermore, we expect that each classifier relies on some arbitrary features. When we are unable to remove the arbitrariness, we may at least reduce the impact of any particular arbitrary feature.

In this example, because systemic exclusion is of moral concern and the loss to the employer or lender from introducing this method is either nonexistent or small, we recommend introducing randomization.[25] The case for randomization is stronger still in contexts in which the algorithm relies on many arbitrary features and does not do a good job of achieving its aims. Rather than leaving out only one or two of the qualified applicants, imagine a screening tool that identifies only two or three qualified applicants among the ten it puts forward. While clearly it would be good for this algorithm to be improved, if errors are to remain it is better for these errors to be distributed rather than concentrated. Introducing randomization achieves this aim as well. In saying so, we are not endorsing the use of poor selection devices. The lender or employer itself will surely be motivated to improve its screening tool to achieve its ends. Rather the point is that at any given time, the employer or lender is likely to be using a tool that it believes is good. If the employer or lender is correct about this, introducing randomness helps distribute opportunity without loss to the employer or lender in most cases, given multiplicity. And if the employer or lender is mistaken

[25]Decision theorists call the choice to introduce randomness to a choice between options, as we do here, a mixed strategy. For more on the normative status of mixed strategies such as these, see Zollman (2019).

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

and the tool it is using is flawed and relies on many arbitrary features, introducing randomness will distribute that burden in a way that prevents systemic exclusion.

5. Conclusion

To conclude, we suggest one additional reason to encourage the use of random classifier ensembles and other techniques meant to introduce randomness to the realms of algorithmic decision-making for hiring and lending. That is that using automated decision-making systems in both realms risks implying more precision than the subject matter of hiring or lending affords.

There are many ways to contribute to a team or an organization and therefore many kinds of good employees. Not all the factors that make an employee good are likely to be measured correctly or represented in the input data, and thus they will not be part of the optimization. Optimizing too far risks over-reliance on data known to be limited.

Randomization addresses the harms of systemic exclusion when it is arbitrary and nonarbitrary. It allows many firms to run the same algorithm on the same applicants but (sometimes) get different results. It thus addresses both the moral and practical problems we raise and constitutes an improvement on the current state of the field.

Acknowledgments. The authors are grateful for helpful written comments from Liam Kofi Bright, Roger Creel, Seth Lazar, Chad Lee-Stronach, Elinor and Kathleen P. Nichols, Kate Vredenburgh, and an anonymous reviewer at the *Canadian Journal of Philosophy* as well as helpful discussion with audiences at TU Eindhoven, Northeastern, Notre Dame, MIT, the PAIS workshop, the Institute for Human-Centered Artificial Intelligence (HAI) at Stanford, the Surrey Centre for Law & Philosophy's Symposium on Ethics and Algorithms, the Workshop on Algorithmic Fairness at Copenhagen, the Algorithms at Work group at Oxford, and the Wharton Legal Studies and Business Ethics Seminar Series at the University of Pennsylvania.

Kathleen Creel is an Embedded EthiCS Postdoctoral Fellow at the McCoy Family Center for Ethics in Society, the Institute for Human-Centered Artificial Intelligence, and the Department of Philosophy, Stanford University.

Deborah Hellman is a professor of law and director of the Center for Law and Philosophy at the University of Virginia School of Law.

References

Albright, Alex. 2019. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." Harvard John M. Olin Fellow's Discussion Paper.

Anderson, Elizabeth S. 1999. "What Is the Point of Equality?" *Ethics* 109 (2): 287–337.

Barocas, Solon, and Karen Levy. 2020. "Privacy Dependencies." *Washington Law Review* 95 (2): 555.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Bright, Liam Kofi, Daniel Malinsky, and Morgan Thompson. 2016. "Causally Interpreting Intersectionality Theory." *Philosophy of Science* 83 (1): 60–81. https://doi.org/10.1086/684173.

Carbado, Devon W. 2013. "Colorblind Intersectionality." *Signs: Journal of Women in Culture and Society* 38 (4): 811–45. https://doi.org/10.1086/669666.

Chemerinsky, Erwin. 1983. "In Defense of Equality: A Reply to Professor Westen." *Michigan Law Review* 81 (3): 575. https://doi.org/10.2307/1288510.

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5 (2): 153–63. https://doi.org/10.1089/big.2016.0047.

Citron, Danielle Keats, and Frank Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89: 1.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." ArXiv:1701.08230.

Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87 (4): 568–89.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Ferguson, Andrew Guthrie. 2017. *The Rise of Big Data Policing Surveillance, Race, and the Future of Law Enforcement*. New York: New York University Press.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

Fishkin, Joseph. 2014. *Bottlenecks: A New Theory of Equal Opportunity*. Oxford: Oxford University Press.
Fourcade, Marion, and Kieran Healy. 2013. "Classification Situations: Life-Chances in the Neoliberal Era." *Accounting, Organizations and Society* 38 (8): 559–72. https://doi.org/10.1016/j.aos.2013.11.002.
Gandy, Jr., Oscar H. 2020. "Panopticons and Leviathans: Oscar H. Gandy, Jr. on Algorithmic Life." *Logic* 12. https://logicmag.io/commons/panopticons-and-leviathans-oscar-h-gandy-jr-on-algorithmic-life/.
Gandy, Jr., Oscar H. 2021. *The Panoptic Sort: A Political Economy of Personal Information*. New York: Oxford University Press.
Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning." ArXiv:1806.00069.
Goh, Gabriel. 2019. "A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features': Two Examples of Useful, Non-Robust Features." *Distill* 4 (8). e00019.3. https://doi.org/10.23915/distill.00019.3.
Green, Anthony. 2021. "Want a Job? The AI Will See You Now." July 7, 2021. *MIT Technology Review* (podcast). https://www.technologyreview.com/2021/07/07/1043089/podcast-want-a-job-the-ai-will-see-you-now-2/.
Greenawalt, Kent. 1983. "How Empty Is the Idea of Equality." *Columbia Law Review* 83. https://scholarship.law.columbia.edu/faculty_scholarship/82.
Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2017. "On Fairness, Diversity and Randomness in Algorithmic Decision Making." ArXiv:1706.10208.
Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." ArXiv:1610.02413.
Hellman, Deborah. 2008. *When Is Discrimination Wrong?* Cambridge, MA: Harvard University Press.
Hellman, Deborah. 2020. "Measuring Algorithmic Fairness." *Virginia Law Review* 106 (4). https://www.virginialawreview.org/articles/measuring-algorithmic-fairness/.
Hendrycks, Dan, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2020. "Natural Adversarial Examples." ArXiv:1907.07174.
Hu, Lily, and Issa Kohler-Hausmann. 2020. "What's Sex Got to Do with Machine Learning?" In FAT* '20: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 513. https://doi.org/10.1145/3351095.3375674.
Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." ArXiv:1905.02175.
Kaushal, Amit, Russ Altman, and Curt Langlotz. 2020. "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms." *Journal of American Medical Association* 324 (12): 1212–13. https://doi.org/10.1001/jama.2020.12067.
Kearns, Michael, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." In PMLR: *Proceedings of the 35th International Conference on Machine Learning* 80: 2564–72. http://proceedings.mlr.press/v80/kearns18a.html.
Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." ArXiv:1609.05807.
König, Pascal D. 2020. "Dissecting the Algorithmic Leviathan: On the Socio-Political Anatomy of Algorithmic Governance." *Philosophy & Technology* 33 (3): 467–85. https://doi.org/10.1007/s13347-019-00363-w.
Marx, Charles T., Flavio du Pin Calmon, and Berk Ustun. 2020. "Predictive Multiplicity in Classification." ArXiv:1909.06677.
Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. "A Survey on Bias and Fairness in Machine Learning." ArXiv:1908.09635.
Mulgan, Tim. 2001. *The Demands of Consequentialism*. New York: Oxford University Press.
Nussbaum, Martha. 2011. *Creating Capabilities*. Cambridge, MA: Harvard University Press.
Peters, Christopher. 1996. "Foolish Consistency: On Equality, Integrity, and Justice in Stare Decisis." *Yale Law Journal* 105 (8). https://digitalcommons.law.yale.edu/ylj/vol105/iss8/1.
Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–81.
Raji, Inioluwa Deborah, and Joy Buolamwini. 2019. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." In AIES: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–35. https://doi.org/10.1145/3306618.3314244.
Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Raz, Joseph. 1988. *The Morality of Freedom. The Morality of Freedom*. Oxford: Oxford University Press.
Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." In KDD: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. https://doi.org/10.1145/2939672.2939778.
Safiya, Noble. 2018. *Algorithms of Oppression*. New York: New York University Press.
Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards. 2020. "What Does It Mean to 'Solve' the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems." In FAT* '20: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–68. https://doi.org/10.1145/3351095.3372849.
Schauer, Frederick. 2003. *Profiles Probabilities and Stereotypes*. Cambridge, MA: The Belknap Press of Harvard University Press.

https://doi.org/10.1017/can.2022.3 Published online by Cambridge University Press

Sen, Amartya. 1980. "Equality of What?" In *Tanner Lectures on Human Values*, 197–220. Cambridge: Cambridge University Press.

Sen, Amartya. 1993. "Capability and Well-Being." In *The Quality of Life*, edited by Martha Nussbaum and Amartya Sen, 30–53. Oxford: Clarendon Press.

Vredenburgh, Kate. 2021. "The Right to Explanation." *The Journal of Political Philosophy* 0 (0): 1–21. https://doi.org/10.1111/jopp.12262.

Vredenburgh, Kate. Forthcoming. "Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace." *Canadian Journal of Philosophy*.

Waldron, Jeremy, and Peter Westen. 1991. "The Substance of Equality." *Michigan Law Review* 89 (6): 1350. https://doi.org/10.2307/1289475.

Westen, Peter. 1982. "The Empty Idea of Equality." *Harvard Law Review* 95 (3): 537–96. https://doi.org/10.2307/1340593.

Zech, John R., Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study." *PLoS Medicine* 15 (11): e1002683.

Zednik, Carlos. 2019. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy & Technology* 34: 254–88. https://doi.org/10.1007/s13347-019-00382-7.

Zollman, Kevin J. S. 2019. "On the Normative Status of Mixed Strategies." Preprint. http://philsci-archive.pitt.edu/17979/.

Cite this article: Creel, K. and Hellman, D. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy* 52: 26–43, doi:10.1017/can.2022.3