Philosophical Studies
https://doi.org/10.1007/s11098-023-02013-6

What we owe to decision-subjects: beyond transparency and explanation in automated decision-making

David Gray Grant[1,2] · Jeff Behrends[3] · John Basl[4]

Accepted: 7 July 2023

Abstract

The ongoing explosion of interest in artificial intelligence is fueled in part by recently developed techniques in machine learning. Those techniques allow automated systems to process huge amounts of data, utilizing mathematical methods that depart from traditional statistical approaches, and resulting in impressive advancements in our ability to make predictions and uncover correlations across a host of interesting domains. But as is now widely discussed, the way that those systems arrive at their outputs is often opaque, even to the experts who design and deploy them. Is it morally problematic to make use of opaque automated methods when making high-stakes decisions, like whether to issue a loan to an applicant, or whether to approve a parole request? Many scholars answer in the affirmative. However, there is no widely accepted explanation for why transparent systems are morally preferable to opaque systems. We argue that the use of automated decision-making systems sometimes violates duties of consideration that are owed by decision-makers to decision-subjects, duties that are both epistemic and practical in character. Violations of that kind generate a weighty consideration against the use of opaque decision systems. In the course of defending our approach, we show that it is able to address three major challenges sometimes leveled against attempts to defend the moral import of transparency in automated decision-making.

Keywords Artificial intelligence · Machine learning · Transparency · Interpretability · Opacity · Decision making · Explanation · Right to explanation

1 Introduction

Institutions increasingly rely on artificial intelligence to help them make high-stakes decisions about how to treat decision-subjects, such as decisions about whom to employ, whose mortgage applications to approve, whom to arrest or imprison, and whom to offer potentially life-saving medical interventions. This trend has been driven in part by the development of powerful new machine learning techniques

Extended author information available on the last page of the article

Published online: 17 August 2023

such as deep learning. While systems based on these techniques promise to make decision-making more accurate and efficient, they are often “black boxes,” the inner workings of which are mysterious even to experts (Breiman, 2001; Burrell, 2016; Doshi-Velez & Kim, 2017). More traditional automated decision systems, by contrast, are “interpretable,” meaning roughly that human experts can explain why they produce the outputs they do by inspecting their underlying mathematical models.

In this paper, we will defend the *Explainability Thesis*, a principle concerning the use of black box systems:

> *Explainability Thesis.* In many contexts, decision-makers are morally obligated to avoid basing their decisions about how to treat decision-subjects on the outputs of black box AI systems.

The Explainability Thesis has broad appeal. Numerous authors have suggested that there is something morally problematic about using black box AI systems (hereafter just “black box systems”) to allocate important benefits and burdens.[1] The idea is also enshrined in many “codes of ethics” for the development of AI-based systems.[2] Some researchers go so far as to say that we ought not use black box systems to make high-stakes decisions at all, at least in cases where there are more explainable alternatives.[3]

However, proponents of the Explainability Thesis face an important challenge. It seems implausible that decision-makers have a *sui generis* duty to avoid relying on black box systems. Insofar as the Explainability Thesis picks out a genuine moral duty, then, it must be grounded in other duties that decision-makers have. But what might those duties be, and how do they give rise to a duty to avoid relying on black box systems in the relevant contexts? One strategy for defending the Explainability Thesis attempts to ground it in what we call *duties of transparency*—duties to disclose information about how the decision-making process works to other parties (Selbst & Barocas, 2018; Vredenburgh 2022). Insofar as even experts are unable to explain the input/output behavior of black box systems, a requirement to disclose meaningful information about a decision-making process will apparently require avoiding the use of such systems, thus vindicating the Explainability Thesis. As we will argue below, however, this strategy for defending the thesis has difficulty explaining its appeal outside of special cases. Call the problem of specifying the nature of the duties that ground a duty to eschew black box systems the *Grounding Problem*.

In this paper, we develop an alternative defense of the Explainability Thesis that appeals to the duty to show *due consideration* to decision-subjects. Decision-makers show due consideration to decision-subjects when they are appropriately sensitive to their moral claims—and more specifically, the moral claims they have that bear on how decisions affecting them should be made. We will argue that basing decisions on the outputs of black box systems is morally problematic in many contexts

[1] Mittelstadt et al. (2016), Floridi et al. (2018) and Rudin (2019).

[2] Mittelstadt et al. (2016), Floridi et al. (2018) and Basl and Sandler (2021).

[3] Rudin (2019).

because doing so interferes with decision-makers' ability to show due consideration to decision-subjects.

Our approach to defending the Explainability Thesis helps us to resolve two additional problems raised by skeptics of the thesis. The *Definition Problem* challenges defenders of the thesis to provide a clear account of what it is for an AI system to be a "black box," as is seemingly needed to address concerns that the concept is not well-defined. The *Double Standard Problem* challenges them to provide a defense of the thesis that does not overgeneralize, condemning decision-making practices that most find unobjectionable, such as basing decisions on the judgment of human experts.

Our plan for the paper is as follows. In Sect. 2 we briefly criticize transparency-centric defenses of the Explainability Thesis. In Sect. 3 we suggest an alternative defense of the thesis, one grounded in our duty to give due consideration to those about whom we make decisions. In Sect. 4 we address the Definition Problem by specifying the class of systems our arguments will target. In Sect. 5 we explain the components of due consideration that have a distinctively *epistemic* character and show how these duties may limit the permissible use of black box systems in decision-making. In Sect. 6 we address the Double Standard Problem. Section 7 considers the components of due consideration that have a distinctively *practical* character, and Sect. 8 offers concluding remarks.

2 The transparency defense

According to the *Transparency Defense*, using black box systems to make high-stakes decisions is problematic in many contexts because decision-makers have *duties of transparency*—duties to disclose certain details about how decisions are made to decision-subjects (or perhaps their advocates, such as third-party watchdogs).[4] To be successful in any given case, the Transparency Defense must establish that two conditions are satisfied: (1) that there is an applicable duty of transparency; and (2) that it requires disclosing information that would not be available if a black box system were used.

[4] See e.g. Selbst and Barocas (2018) and Vredenburgh (2022). Note that proponents of the Transparency Defense typically take the relevant duties of transparency to be derived; the idea is that disclosing information about the decision-making process to decision-subjects or their representatives helps decision-makers satisfy more foundational duties to decision-subjects (which could potentially be satisfied through other means). For example, Selbst and Barocas identify three morally important goals that such disclosures advance, and that ground "popular and scholarly calls for explanation": (1) respecting the "personhood" of decision-subjects; (2) enabling decision-subjects to "effectively navigate the decision-making process"; and (3) facilitating assessment of the fairness and legality of the decision-making process by decision-subjects and their representatives (1118–1126). For their part, Selbst and Barocas argue that such calls are better served by disclosing information about how the model is used in decision-making as well as "the institutional and subjective process behind its development" than information about how the model itself works (1130). Vredenburgh argues that decision-subjects have a "right to explanation" that is grounded in (2) and (3) above, and that respecting this right requires eschewing models that cannot be made "functionally transparent" in Creel (2020)'s sense.

Both claims seem plausible in at least some contexts. For example, US law requires lenders denying credit to provide the applicant with an easy-to-understand explanation of which features of their application played the biggest role in the decision. The rationale for this is that the explanations make it easier for decision-subjects to contest inaccurate or illegal decisions, as well as to determine how they can achieve better results in future interactions with the credit system. Reliance on a black box system such as a deep neural network (DNN) would make it difficult or impossible to provide this information, and lenders typically use interpretable models instead (Selbst & Barocas, 2018).

However, it is not clear that these two conditions are met outside of special cases. Regarding the first condition, many high-stakes decisions are not obviously governed by duties of transparency. Employers, for example, are not legally required to explain the underlying logic of their hiring decisions to applicants, and arguably are not morally obligated to do so, either.[5] Regarding the second condition, there are cases where duties of transparency apply, but relying on a black box system seems consistent with satisfying those duties. For example, London (2019) points out that doctors are very often incapable of explaining how the methods they rely on to make diagnoses work because the relevant mechanisms are not well-understood. However, London argues that this does not prevent doctors from meeting their duties of transparency to patients, because they have no obligation to disclose that sort of information.

Duties of transparency, however, are not the only duties that we have to decision-subjects. There are, in addition to whatever reasons might be present for explaining our decisions to others, moral constraints on how we make those decisions in the first place. For example, a judge who decides whether to grant bail in a pretrial hearing by flipping a coin wrongs the defendant in question, even if she freely discloses how she made her decision. Intuitively, to make such an important decision in such an arbitrary way is to fail to show *due consideration* to the defendant, who has important rights and interests at stake in the decision that the judge is obligated to respect.

We will argue that the obligation to show due consideration provides decision-makers with strong (but potentially overridable) reason to avoid relying on black box decision systems in a wide range of contexts. We begin by explaining what due consideration is in more detail.

3 Due consideration

Developing a complete theory of due consideration is beyond the scope of this paper, but we can identify the broad outlines of one, and attempt to show how we can make progress toward defending the Explainability Thesis armed only with those theoretical contours.[6] In our view, a decision-maker D shows due consideration to

[5] See Vredenburgh (2022) for an argument that employers do have this moral obligation.

[6] See Scanlon (2018) on the need for a theory of due consideration, as well as for a recent discussion of the concept and its relationship to procedural fairness.

decision-subject S just in case D adopts decision procedures that are appropriately responsive to S's moral claims on the decision process—claims that S has that place restrictions on how D ought to make decisions about how to treat S.

The duty to show due consideration can be decomposed into a variety of constituent duties that are grounded in different kinds of moral claims that decision-subjects have. We call these *duties of consideration*.[7] In the remainder of this section, we will distinguish different types of duties of consideration in order to flesh out our account of due consideration and lay the groundwork for the ensuing discussion.

First, we can distinguish between *substantive* and *procedural* duties of consideration, which are grounded in different kinds of claims decision-subjects can have on how a decision procedure works: substantive and procedural. By "substantive claims," we mean claims to be treated in certain ways in virtue of the features that the decision subject in fact has (as opposed to features the available evidence suggests they have). For example, an innocent defendant in a criminal trial has a substantive claim to be found innocent and released. Such features are often not directly perceptible, but instead need to be inferred. In such cases, there will normally be some risk that decision-makers will make incorrect inferences and thus fail to treat the decision subject in the way that they are substantively owed.[8] Procedural fairness requires that these risks be managed using appropriate procedural safeguards (such as competent legal representation).[9] A decision procedure that fails to provide appropriate safeguards—thereby exposing decision-subjects to an excessively high risk of being treated in substantively unfair ways (such as wrongful imprisonment)—is procedurally unfair in virtue of failing to show due consideration to those subject to it. Procedural claims, by contrast, are claims constraining the set of permissible decision procedures that are not grounded in decision-subjects' substantive claims. For example, suppose a prosecutor seeks to use information obtained from an illegal wiretap against a criminal defendant that prosecutors know is guilty. This is procedurally unfair, but presumably not in virtue of the defendant's substantive claim against wrongful conviction.

The distinction between substantive and procedural duties of consideration crosscuts another distinction that will be important for our purposes. Deciding how to treat others requires performing two different tasks: (a) gathering and evaluating evidence to form beliefs about what decision-subjects are like in morally relevant respects; and (b) deciding how to treat them given those beliefs. The first task, *fact-finding*, is epistemic (or zetetic) in nature; the second task, *decision-making*, is a practical reasoning task. What we call *duties of evidential consideration* constrain how fact-finding is conducted. What we call *duties of practical consideration* constrain decision-making.[10] Broadly speaking, duties of evidential consideration apply to how decision-makers answer descriptive questions about decision-subjects,

[7] The duty to show due consideration thus unifies an important part of what Enoch (2018) calls "evidence law for morality.".

[8] Rawls (1999) calls these cases of "imperfect procedural justice."

[9] Di Bello and O'Neil (2020) call this requirement "due concern" in the special case of criminal trials.

[10] We are grateful to Stephanie Sheintul for showing us that the terminology we were using for these duties in an earlier draft of this paper was potentially confusing to readers.

whereas duties of practical consideration apply to how they answer normative questions about them (in particular, questions about how they ought to be treated).

With the nature of our solution to the Grounding Problem on the table, we are now in a position to tackle the Definition Problem by specifying the class of systems our arguments will target.

4 The definition problem

The concept of a "black box" AI system is often defined in contrast to "explainable" or "interpretable" AI systems. However, the literature on explainable artificial intelligence (XAI) often emphasizes that these concepts lack agreed-upon definitions, and pick out a variety of seemingly disparate properties (Lipton, 2018).[11] Skeptics of the Explainability Thesis contend that, because explainability and interpretability remain poorly understood, claims about their moral significance are difficult to evaluate, or even to formulate in a suitably rigorous way. Krishnan (2019) finds it "worrying," for instance, "that so much importance has been afforded to interpretation in the absence of an adequate grasp of what the concept means when applied to algorithms."

To address this worry, defenders of the Explainability Thesis need to say more precisely what they mean by "black box system." We will define it in terms of three concepts: flexibility, dimensionality, and rule transparency.

The systems that are collectively referred to as "black box AI" share two technical properties. On the one hand, they are highly *flexible*, which means that they are capable of modeling a much broader range of relationships between inputs and outputs than, say, linear models are (James et al., 2021). They are also highly *dimensional*, in the sense that they perform computations over very many input features (Selbst & Barocas, 2018). In combination, these two properties contribute significantly to both the power of contemporary black box AI systems and their tendency to resist explanation (Breiman, 2001; Selbst & Barocas, 2018). For example, DNNs can be trained to compute a vast array of complex, nonlinear mathematical functions over a vast number of datapoints about a data-subject. This high flexibility and dimensionality helps to explain why they often can make more accurate predictions than simpler predictive models—because the world is often complicated, and they can capture more of that complexity—but it also means that it is in general difficult to explain a DNN's predictions in terms that humans are capable of understanding.[12]

This brings us to what we will call "rule transparency." Rule transparency is a species of what Creel calls "functional transparency." A system is *functionally transparent* for some agent to the extent that the agent is in a position to know what

[11] The terminological situation here is indeed vexed (Clinciu and Hastie, 2019), with some researchers contrasting black box systems with "interpretable" systems (Burrell ,2016; Krishnan ,2019), others with "explainable" systems (Speith, 2022; Baum, 2022), and others still with "scrutable" systems (Selbst and Barocas, 2018).

[12] Barocas and Selbst, (2018). There is no guarantee that a DNN's behavior is even amenable to explanation in human-understandable terms—the decision rules it implements may defy compact summary, or refer to features that are too "alien" for humans to understand (Buckner, 2020; Creel ,2020).

higher-level computations the system performs in order to transform inputs into outputs (Creel, 2020). Our notion of rule transparency is defined in terms of two kinds of higher-level computations, those that apply inference and decision rules. An *inference rule* is any rule used to answer descriptive questions about decision-subjects, and a *decision rule* is any rule used to decide how to treat particular decision-subjects, given their descriptive properties. Say that a system *implements* an inference or decision rule when it is disposed to behave in ways that can be accurately explained in terms of its applying the rule to decision-subjects. The inference and decision rules implemented by a system thus constitute what is sometimes called its "decision logic." Finally, say that a system is *rule transparent* to an agent to the extent that the agent is in a position to know what inference and decision rules it implements.

Computer scientists distinguish between global explainability and local explainability.[13] Global explainability has to do with agents' ability to provide unified explanations of a system's decision-making behavior across a broad range of background conditions, whereas local explainability has to do with their ability to explain the system's behavior on particular occasions.[14] Rule transparency has both global and local aspects. What we will call a system's *global rules* allow us to provide unified explanations of its behavior across a broad range of decision-making situations, whereas its *local rules* allow us to explain its behavior in special situations where its global rules fall short. A system is rule transparent, in our sense, to the extent that both its global and local rules are known.

Knowing a model's global rules is insufficient for assessing whether using it would be consistent with due consideration. To see this, suppose that a machine learning system used by courts to assess recidivism risk implements a rule that anyone named "Aloysius" is to be treated as at extremely high risk of recidivism, regardless of what other evidence of risk is available. Suppose also that the name is so uncommon that the rule will virtually never come into effect, suggesting that the rule cannot be counted among the system's global rules. Decision-makers still have a pro tanto reason of due consideration not to use the system, in light of the fact that it implements this local rule. While we will suppress this complication in what follows, our arguments below suggest that decision-makers' ignorance of either global or local rules used by a system can interfere with their ability to show due consideration.[15,16]

[13] See e.g. Doshi-Velez and Kim (2017) and Speith (2022).

[14] Note that this is our rational reconstruction of the distinction, not a canonical definition.

[15] For example, it is tempting to assume that the overall accuracy of a system (Sect. 5.1) will depend almost exclusively on its global rules. However, a black box system might rely heavily on local rules tailored to highly specific situations, in which case its accuracy would depend heavily on its local rules. Similarly, a system might exploit morally inadmissible evidence (Sect. 5.3) in only its global rules, only its local rules, or some combination of the two. One upshot of this is that our account may support more stringent explainability requirements in some contexts than Vredenburgh (2022)'s, as Vredenburgh argues that satisfying the "right to explanation" she articulates requires only knowledge of a system's global rules.

[16] Note that Langer et al. (2021) argue, in a similar vein, that both global and local explainability can be instrumental for ensuring that a decision system "complies with ethical standards" (p. 8).

We are now in a position to sharpen up our version of the Explainability Thesis in a way that addresses the Definition Problem. We will use "black box system" to refer to AI systems with the following three features: (1) high flexibility; (2) high dimensionality; and (3) limited rule transparency. We intend "high" and "limited" here to be interpreted in such a way that our definition of "black box system" picks out roughly the class of systems that AI researchers currently have in mind when they talk about "black box AI," such as those based on deep neural networks and random forests. Black box systems in this sense contrast with so-called "interpretable" systems, which are inherently much less flexible and high-dimensional, but also more rule transparent.[17]

Two clarifications.

First, we concede that the resulting boundary between "black box" and "interpretable" systems is not sharp. However, a rough-and-ready characterization of the class of systems our arguments target will suffice for our purposes. Our goal is to clearly articulate one class of concerns about the use of black box AI to make high-stakes decisions, and to shed light on how to analyze and address these concerns. We concede that judgment will be required to determine how our arguments apply to any *particular* automated decision system. (Bear in mind that our claim is that decision-makers *often* have an obligation not to rely on black box systems, not that they *always* do.)

Second, whether a particular system is a black box in our sense is subject to change as the result of empirical efforts to increase the system's rule transparency. We concede that it may be possible, in some cases, to render a black box system sufficiently rule transparent to neutralize our concerns.[18] Our argument applies solely to systems for which such efforts have not yet succeeded.

This concludes our discussion of the Definition Problem. Our goal for the rest of the paper is to defend the Explainability Thesis by showing how, in a broad variety of contexts, relying on black box systems can interfere with decision-makers' ability to discharge their duties of consideration. At a high level, we will identify two different kinds of interference. First, some duties of consideration enjoin decision-makers to adopt a decision-making procedure that implements inference or decision rules that satisfy particular constraints: constraints on the content of the rules or their likely effects. Relying on a black box system will often interfere with their ability to ensure, to an adequate degree, that these constraints are satisfied. Second, some duties of consideration require the practical reasoning component of decision-making to be delegated to full-blown moral agents exercising their moral reasoning capacities, rather than automated systems without these capacities.

[17] On this usage of "interpretable," see Rudin (2019) and Bell et al. (2022).

[18] The goal of XAI research is to develop tools that can be used to enhance the explainability of black box AI systems. However, existing tools have important limitations (Zerilli, 2022; Creel ,2020; Fleischer, 2022) and require significant resources to deploy effectively. In cases where decision-makers *could* render a black box system rule transparent through the use of these tools, our claim is that they are either obligated to do so or to avoid using the system. See Minh et al. (2022) for a recent overview of the XAI literature, and Langer et al. (2021) for discussion of how XAI tools can be used to address specific concerns about black box systems violating moral standards.

The next section will focus on the first type of interference as it applies to duties of evidential consideration. We will then discuss both types of interference as they apply to duties of practical consideration.

5 Duties of evidential consideration

Decision-makers often have duties of evidential consideration not to base fact-finding on the outputs of black box systems. At first glance, this claim might seem surprising. After all, the standard line on these systems is that they make more accurate predictions than is possible using more traditional (and rule transparent) methods.[19] Indeed, black box systems are responsible for some of the most impressive success stories of contemporary artificial intelligence research, such as systems for predicting breast cancer from mammograms that outperform human radiologists (McKinney et al., 2020). Further, there is a large body of research showing that actuarial methods for making predictions outperform those that rely on the clinical judgment of human experts across a broad range of tasks and domains.[20] If actuarial methods are more accurate than those that rely on human judgment, and black box systems are the most accurate actuarial methods available, then it might seem that using black box systems is the best way for decision-makers to discharge their duty to form beliefs about decision-subjects in a way that is appropriately sensitive to their claims. Call this *the argument from accuracy*.

The argument from accuracy may seem compelling. However, there are several ways in which relying on black box systems can lead decision-makers to fall short in terms of evidential consideration. First, black box machine systems are not *always* more accurate than traditional predictive methods, and it can be difficult to anticipate whether they will maintain the high level of accuracy exhibited on sample data when they are deployed in the field. Second, black box systems have a tendency, when compared to human decision-makers, to ignore relevant evidence that they have not specifically been designed to take into account. And third, black box machine learning systems can, unbeknownst to their designers, rely on morally inadmissible evidence—evidence that decision-makers have an obligation to set aside.

We elaborate on each of those points in the three following subsections, beginning with a closer look at the argument from accuracy.

5.1 Accuracy

Consider the practice of detaining criminal defendants pretrial on the basis of estimated recidivism risk. The justification for the practice goes something like this. Preventive detention is *substantively* fair in cases where the defendant poses a sufficiently great danger to the public to outweigh their claim against detention. Preventive detention is *procedurally* fair when the available evidence supports the

[19] See e.g. Breiman et al. (2001); Caruana et al. (2015).

[20] For a list of key references, see the introduction to Jung et al. (2020).

conclusion that the defendant poses a sufficient danger to make preventive detention substantively fair. Further, such evidence is often available in particular cases, and courts are competent to evaluate that evidence. Therefore, the practice of pretrial detention on the basis of estimated recidivism risk is procedurally fair, at least in principle.[21]

Here's how the argument from accuracy applies to this example. Showing due consideration to defendants in pretrial hearings requires being appropriately sensitive to the claims they have that bear on how they ought to be treated by the state. In particular, defendants that pose a low risk of recidivism have substantive claims against detention; courts therefore have substantive duties of consideration to be appropriately sensitive to those claims. Sensitivity to substantive claims is a matter of predictive accuracy; therefore, courts are sensitive to defendant's substantive claims to the extent that they use accurate methods to estimate recidivism risk. Since using a black box system is typically the most accurate predictive method available, we should expect estimating recidivism risk using a black box system to be the best way to show due consideration to defendants.

One thing this argument gets right is that the (relative) accuracy of a predictive method does make a difference to whether using it would be consistent with showing due consideration to decision-subjects.[22] Suppose that the only available way to estimate recidivism risk is by using one of two algorithms. One is COMPAS, a recidivism prediction algorithm used by courts across the US. Assume that COMPAS is known to be highly accurate (by any standard measure). The other is TEALEAVES, whose scores are randomly generated and provide no information about recidivism risk. Suppose courts know all this, but use TEALEAVES to make pretrial detention decisions anyway. Further, suppose that Sacco and Vanzetti are wrongly accused of murder, and that neither poses any danger to others. Sacco receives a high TEALEAVES score and is detained pretrial on that basis; Vanzetti receives a low score and is released. Sacco has two substantive claims against being treated in this way. On the one hand, he has a noncomparative claim against detention, since by stipulation he is insufficiently dangerous to make detention substantively fair. On the other hand, he has a comparative claim against being treated less favorably than Vanzetti, as there is no basis for treating him less favorably. Sacco's treatment is thus substantively unfair on both comparative and noncomparative grounds.[23] Moreover, by knowingly using an inaccurate method to estimate recidivism risk when an

[21] We take no stance on whether preventive detention is justifiable. See Mayson (2018) for discussion.

[22] Our discussion here focuses on overall accuracy, rather than accuracy for particular subpopulations. However, we should note that decision-subjects may also have claims against the use of fact-finding methods that are *differentially* accurate for different groups of decision-subjects. For example, many popular "statistical criteria of fairness" enjoin decision-makers to use predictive methods that achieve similar levels of accuracy—measured in one way or another—across different social groups (Corbett-Davies et al., 2023). For example, Equalized Odds requires that the false-positive and false-negative rates of a fact-finding method be the same for each social group (Hardt et al., 2016; Castro, 2019; Grant, 2023). Since satisfying these criteria can require sacrificing overall accuracy, computer scientists sometimes speak of an "accuracy-fairness tradeoff" (Dutta et al., 2020; Rodolfa et al., 2021). Our arguments here show that this locution is misleading, insofar as overall accuracy and accuracy across subpopulations both matter to fairness.

[23] On the distinction between comparative and noncomparative fairness, see Feinberg (1974).

accurate one was available, the court has failed to be appropriately sensitive to both of these two claims—and so has violated its duties of evidential consideration to Sacco.

We concede, then, that decision-makers often have reason to believe that using a black box system would allow them to be more sensitive to decision-subjects' substantive claims than the available alternatives. And we concede that this gives them a reason to think that using a black box system would help them show due consideration in fact-finding. However, the argument from accuracy faces two important objections.

First, experts often suggest that relying on a black box system in high-stakes contexts is problematic on the grounds that these systems may not perform nearly as well in the field as they do in the lab. This tendency results from three features that they share. First, black box systems characteristically require significantly more input data about decision-subjects than interpretable ones, which raises the likelihood that transcription errors and other data quality issues will lead to inaccurate predictions (Rudin, 2019). Second, as discussed above, black box systems are based on highly flexible machine learning techniques, which means that they are capable of modeling a much broader range of relationships between inputs and outputs than, say, linear models are. This can help them achieve impressive gains in accuracy, but it also makes them more vulnerable to *overfitting*, which occurs when a predictive model incorrectly generalizes from idiosyncratic patterns in the training data, patterns that are unlikely to be present in the context of deployment.[24] For instance, one resume screening tool based on machine learning "learned" that applicants who were named "Jared" were more likely to be strong performers, presumably because the company that provided the training data once had a star employee named "Jared" (Shellenbarger, 2019). Third, since the inference rules a black box system implements are not known, decision-makers will be in a worse position to detect cases of overfitting than when rule transparent systems are used.[25] This has led some researchers to conclude that black box systems should not be used in high-stakes applications, such as health care.[26]

To summarize, black box systems share features—their vulnerability to data quality problems, their tendency to overfit their training data, and their lack of rule transparency—that create a risk that their performance in the field may be far worse than their performance during development. If an interpretable model is available, then we have a reason grounded in the duty of evidential consideration to prefer it, even if pre-deployment testing suggests that it is somewhat (and perhaps even significantly) less accurate in testing conditions.[27]

Second, recall the distinction between substantive and procedural duties of consideration. Substantive duties of consideration enjoin decision-makers to use

[24] On flexibility and its relationship to overfitting, see James et al. (2021).

[25] Caruana (2015), Rudin (2019) and Creel (2020).

[26] See e.g. Caruana et al. (2015).

[27] Recent empirical work suggests that the accuracy gap between black box and interpretable systems—and so the supposed "accuracy-explainability tradeoff"—may be insignificant in many contexts (Bell et al., 2022; Rudin, 2019).

fact-finding methods that are appropriately sensitive to decision-subjects' substantive claims to be treated in particular ways. But it does not necessarily follow that the best way to show due consideration on balance will be to use the most accurate fact-finding methods available, because those using those methods might violate weighty procedural claims that decision-subjects have. Below, we identify two kinds of procedural claims that constrain fact-finding, duties to avoid ignoring readily available evidence, and duties to avoid basing decisions on morally inadmissible evidence.

5.2 Ignoring available evidence

The aforementioned claim that predictive algorithms tend to be more accurate than human decision-makers is a claim about averages—the thought is that, if a well-designed predictive algorithm and a human expert both make a thousand predictions, the algorithm will tend to make fewer mistakes on average than the human expert (at least in many domains). However, predictive algorithms can be completely insensitive to readily available evidence that a human decision-maker would be unlikely to miss, resulting in avoidable mistakes that constitute failures of due consideration. Basing fact-finding on a black box system in particular compounds this risk, since it interferes with decision-makers' ability to determine whether and how particular pieces of evidence are influencing the system's outputs.

To see that predictive algorithms can be insensitive to available evidence that a human decision-maker would not overlook, let's return to the case of COMPAS, and retain our assumption that COMPAS scores are highly accurate at the population level. Suppose that a judge is deciding whether to grant bail to a defendant with an extensive criminal record. Given the defendant's criminal past, he naturally receives a high COMPAS score. However, the judge has an additional piece of evidence, beyond the defendant's COMPAS score. The defendant's neurologist has testified that his past criminal behavior was the result of a brain tumor that has since been successfully excised, and that he now poses a low risk to the public as a result. Since COMPAS was not designed to take evidence of this kind into account, it mistakenly labels the defendant as high risk.

Clearly, it would be unfair for the judge to ignore the neurologists' testimony and refuse to grant bail to the defendant. What this hypothetical example shows is that showing evidential consideration to decision-subjects requires more than making accurate decisions on average. It requires, further, that decision-makers not ignore readily available evidence that would benefit particular decision-subjects. Using a decision procedure that is insensitive to readily available evidence that would benefit some decision-subjects is, therefore, procedurally unfair to those decision subjects.[28,29]

This leads us to a second way in which relying on a black box system can lead to failures of evidential consideration. Black box systems share the general limitation of predictive algorithms that we have just identified: they can respond only to a restricted range of evidence. As we have seen, this means that such systems may fail to take into account readily available evidence that would benefit particular decision subjects. Further, since black box systems are not rule transparent, decision-makers relying on them will have a difficult time determining which pieces of evidence are and are not being taken into account[30]—raising the risk that they will fail to respond appropriately to readily available evidence that would benefit particular decision subjects. This gives decision-makers a second reason of due consideration to avoid relying on black box systems in fact-finding.

5.3 Morally inadmissible evidence

To achieve gains in predictive accuracy, black box systems base their predictions on far more features, and on far more complex relationships among features, than simpler predictive algorithms (Breiman, 2001). This raises the concern that these systems will inadvertently base their predictions on features of individuals that are morally inadmissible as evidence in fact-finding.

A piece of evidence E is *morally inadmissible evidence* for an agent A making a decision D when A is morally obligated to "set aside" E in making D, in the sense that A must reason about what the correct decision for her to make would have been if she had not had E, and decide accordingly.[31] Consider cases of statistical discrimination, in which a decision-maker bases a decision about how to treat a particular person on perceived statistical facts about the group(s) to which they belong. For example, suppose an employer prefers not to hire members of a particular racial group because she believes that they are less qualified on average than members of other groups due to structural discrimination. In this case, an applicant's race is taken to be evidence that they have some further feature, poor future job performance, that is generally taken to be relevant to whether they ought to be hired. Even

[28] This is consistent with the procedure being the best available, all things considered; however, the pro tanto reason against implementing it would remain.

[29] Lippert-Rasmussen (2011) considers and rejects the claim that decision-makers have a duty to consider all available evidence. However, he leaves open the possibility that doing so might be required when there is reason to believe that the relevant evidence would benefit the decision-subject in question, as we argue here. Beeghly (2018) also expresses sympathy for this possibility. Note that decision-makers may also be obligated to give special weight to particular kinds of evidence, such as evidence that is produced by decision-subjects' exercise of autonomy (Eidelson, 2013) or so-called "individualized evidence" (Thomson, 1986). If so, then they have a further reason to avoid basing fact-finding on black box algorithms, since they will be unable to determine if they are giving such evidence appropriate weight. For a recent survey of the (large) literature on the distinction between individualized and "naked" statistical evidence and its moral significance, see Enoch and Spectre (2021).

[30] Kim (2016) makes a similar point.

[31] The term "morally inadmissible evidence" appeared in a draft of Enoch (2016), but not the final version.

if we suppose that the employer is right about the statistical relationship between race and job performance, however, it seems unfair for her to take this evidence into account in making hiring decisions. Instead, the employer is morally obligated to set aside the applicant's race in making her decision, deciding whom to hire as if she did not have this piece of evidence.[32] In other words, the applicant's race is morally inadmissible evidence for purposes of making hiring decisions, at least insofar as taking it into account would disadvantage the applicant.

In contexts in which some evidence is morally inadmissible, decision-makers have an obligation to take reasonable steps to avoid relying on it.[33] In the rest of this section, we argue that the practice of evaluating decision-subjects using black box algorithmic systems often carries a significant risk that morally inadmissible evidence will inadvertently be relied on, and that (as a result) decision-makers often have weighty reason to avoid relying on such systems.

The risk that black box systems will inadvertently exploit morally inadmissible evidence arises from several general features of how they are developed and structured. First, the datasets used to train machine learning systems often encode features that would be morally inadmissible bases for decision-making (such as information about race or gender in the context of hiring or lending). This information may be encoded explicitly or implicitly. For example, information about race or gender is often "redundantly encoded" in the data used by machine learning systems, in the sense that it can be inferred from other features even if explicit references to it have been removed (Dwork et al., 2012).[34] Second, these inadmissible features are often statistically correlated in the training data with the features that decision-makers are trying to predict in the sample data.[35] This may occur either because the correlations are genuine and the training data accurately reflects them, or because the

[32] While the intuition that statistical discrimination is morally wrong in many contexts is widespread, theorists disagree about the best way to diagnose it. See e.g. Bolinger (2021), Eidelson (2015), and Lippert-Rasmussen (2011).

[33] To see this, suppose that you are the CEO of a small company, and are deciding which of two employees to delegate hiring responsibilities to. In the time you have known him, Employee One has made a number of comments that give you some evidence that he might be biased against women candidates. However, the evidence is weak, and you have no other reason to suspect that he would be more biased than anyone else. Employee Two has given you no such cause for concern. In this case, the fact that your evidence suggests that Employee One would be more likely to evaluate women candidates unfairly than Employee Two gives you a reason of due consideration not to delegate hiring responsibilities to Employee One. More generally, if there is a duty to avoid basing decisions on prohibited inference rules, then there is a duty to mitigate the risk that you will rely on such rules.

[34] While discussions of "the problem of redundant encodings" typically focus on protected class membership, the datasets used to train machine learning systems are liable to contain other forms of inadmissible evidence as well. Consider two examples. (1) Basing decisions on features of our personal lives, such as our hobbies and interests, may violate our right to privacy by impinging on our interest in having protected "zones" where we can act freely without fear of being observed (Scanlon 1975). (2) Hellman (2023) argues that we have a duty to avoid *compounding injustice* by basing decisions on features of decision-subjects that are caused by past injustice, and that predictive algorithms based on large datasets are liable to encode such features.

[35] Dwork et al. (2012) and Johnson (2021).

training data is biased in a way that results in spurious correlations.[36] Third, black box systems excel at identifying and exploiting unforeseen statistical correlations in training data, in part due to their high flexibility and dimensionality (as discussed above). This means that these systems will readily exploit inadmissible features if doing so increases performance on training data, as it often does.[37] Fourth, the fact that these systems are not rule transparent entails that it will in general be difficult, if not impossible, for decision-makers to determine whether morally inadmissible evidence is being used.[38]

Putting these together, relying on a black box system will often put decision-makers in a position where (a) there is reason to suspect that the system is basing its predictions on inadmissible features of decision-subjects, but (b) there is no practicable way to determine whether that suspicion is correct.[39] As a result, we argue, decision-makers often have a reason not to base fact-finding on black box machine learning that is grounded in the duty to set aside morally inadmissible evidence.[40]

We anticipate two objections to this line of argument.

First, it might be objected that the prohibition against relying on morally inadmissible evidence is a prohibition against human decision-makers basing their beliefs about decision-subjects on certain kinds of evidence. But neither the human decision-makers nor the system in question here are doing that in the cases just described. On the one hand, the decision-makers are basing their beliefs on facts about the outputs of the system, not on the prohibited facts about decision-subjects. On the other hand, the system doesn't have beliefs in the relevant sense of "belief," and *a fortiori* isn't "basing" its beliefs on inadmissible evidence.[41] Therefore, it may not seem obvious that we have identified a reason to think that basing fact-finding on black box machine learning risks violating the duty to set aside morally inadmissible evidence.

[36] Mayson (2019) and Obermeyer et al. (2019). See also Wachter et al. (2021), especially its related and useful distinction between "bias preserving" and "bias transforming" fairness metrics. While the topic of that paper primarily concerns the relationship between fairness metrics as classified by their proposed distinction, on the one hand, and EU discrimination law on the other, we note that it may prove informative to compare the argumentative project of this paper to Wachter et al.'s observations about the import of the "positive normative choice" confronting decision-makers who opt to use bias transforming metrics in some decision context or other. We are grateful to an anonymous referee for drawing our attention to this work.

[37] Barocas and Selbst (2016) and Johnson (2021).

[38] Kim (2016) and Langer et al. (2021) make similar points.

[39] In some cases it may be practicable, but costly; see footnote 18 above and associated text.

[40] An anonymous reviewer worries that a similar concern can be raised about systems based on interpretable machine learning techniques. In particular, while the features that these models exploit can be freely inspected, those features might be related to other features (e.g. race) in ways that render them morally inadmissible. However, there is still a crucial difference between interpretable and black box models. Decision-makers investigating whether a black box system exploits morally inadmissible evidence must first determine what inference rules it implements and then assess whether those rules operate over morally inadmissible features. Decision-makers evaluating an interpretable model can skip the first step, making their task easier and more likely to be successful.

[41] We concede that acting *as if* these systems have beliefs may be useful for some purposes, as Zerilli et al. (2018) and Zerilli (2022) suggest.

However, the prohibition against relying on morally inadmissible evidence is best understood as a prohibition against using epistemic methods that *implement prohibited inference rules*, regardless of whether those inference rules are implemented by rational agents or computer algorithms. For example, suppose an employer uses a computer program to evaluate employees for merit-based raises that was written by a former employee—an algorithm that, unbeknownst to the employer, explicitly adds points if the employee is a man. If employers have a duty not to base raises on gender, then they presumably also have a duty to avoid using this algorithm, and for the same reasons.

Second, we suggested above that we can't just fix the problem by eliminating references to inadmissible features in the training data, because those features are often "redundantly encoded." Why, though, should we think that a system that exploits features that "redundantly encode" gender (for example) is basing its predictions on gender, as opposed to statistical correlates of gender? Basing decisions on statistical correlates of protected class membership isn't prohibited in general (consider the feature *having a Ph.D. in Philosophy*).

Two responses.

First, in cases where a feature is morally inadmissible, close statistical proxies for it are often inadmissible as well. For example, Amazon was recently forced to mothball a machine learning system it hoped to use to evaluate job candidates after discovering that it had learned to downgrade candidates whose resumes included the word "women's" (as in "women's college" or "women's soccer").[42] Similarly, the prohibition against basing hiring decisions on race plausibly generates a derived duty not to hire on the basis of close proxies for race such as shopping online at certain stores, belonging to certain "cultural affinity" groups on social media, or accessing the internet from certain geographical areas. When relying on a proxy for a feature violates the prohibition against relying on the feature itself is an open question (see Hu forthcoming), but some cases are fairly clear. It is plausible that many datasets will include such features, and decision-makers have a duty to avoid using epistemic methods that exploit them.

Second, the fact that a system's lower-level computations do not operate on explicit representations of prohibited features does not entail that it is not performing such computations at a higher level of abstraction. Many researchers believe that deep neural networks are able to perform tasks such as image recognition because successive layers in the network are able to infer successively more abstract features of the input data (e.g., this is an image of a woman with glasses) (Buckner, 2018, 2019, 8–9). These features need not be represented explicitly by individual nodes or "neurons" in the network, but may instead be represented in a distributed way by groups of nodes working together—just as the neurons in your brain work together to implicitly represent various high-level facts about your environment (see Buckner and Garson 2019, Sect. 6). Therefore, if information about a prohibited feature is redundantly encoded in a black box system's training data, then the system might end up implementing inference rules that *directly* base predictions on that feature, even if the feature is not explicitly encoded in the training data. Since (as noted

[42] Dastin (2018).

above) such information is often useful for making predictions, the risk of this happening may be significant.

6 The double standard problem

Before turning to practical consideration, we should say something about the Double Standard Problem. Psychological research (growing out of Gazzaniga's work with split-brain patients in the 1970s) has cast serious doubt on the idea that we have reliable introspective access to our own decision-making processes.[43] Even if we assume that human decision-makers are in general in a position to know why they decided as they did, they may not be motivated to report their motivations truthfully. Taken together, these considerations suggest that human decision-makers are "black boxes" in the same sense that black box AI systems are. But most defenders of the Explainability Thesis would not want to say that it is morally impermissible to base decisions on human expert judgment! Since defenders of the Explainability Thesis condemn reliance on black box algorithms but not humans, they would seem to be committed to an objectionable double standard (Zerilli et al. 2019).

So far, we have defended the Explainability Thesis in the following way. In many contexts, decision-makers have duties of evidential consideration that require them to adopt a decision procedure that implements inference rules satisfying various constraints, such as that they limit the risk of certain kinds of errors or be sensitive to an appropriately circumscribed range of evidence. Black box systems have a variety of features that make it likely that they will implement inference rules that are prohibited by these constraints. Moreover, since the systems are not rule transparent, it will not in general be practicable for decision-makers to safeguard against this possibility effectively.

This defense appears to run headlong into the Double Standard Problem. After all, aren't human decision-makers prone to implementing prohibited inference rules? And isn't it true that we are not, in general, in a position to tell what inference rules we are implementing? This suggests that relying on human decision-makers *also* carries a significant risk that prohibited inference rules will be implemented, a risk that cannot be controlled adequately due to the black box nature of human decision-making. Consider studies finding that doctors are liable to commit the base rate fallacy when interpreting test results (see e.g. Bramwell et al., 2006). Consider also morally inadmissible evidence: there is considerable evidence that human decision-makers often take social group membership into account (whether consciously or unconsciously) in a way that seems morally wrong.[44] Indeed, one widely cited explanation of the apparent prevalence of "algorithmic bias" is that algorithmic systems are often trained on judgments made by humans, and inherit their biases (Corbett-Davies and Goel, 2018).

So the arguments that we make above seem to generalize to give us reasons against relying on human decision-makers, and not just black box systems. Why,

[43] Schwitzgebel (2019), Sect. 4.2.1.

[44] Bertrand and Mullainathan, (2004), Howell and Korver-Glenn (2018) and Hoffman et al. (2016).

then, don't we say that decision-makers ought to avoid relying on human decision-makers as well? Aren't we guilty of applying an objectionable double standard to humans and machines?

Two responses.

(1) We agree that our argument generalizes to human decision-makers to some extent—just not that it *over*generalizes. Where there are reasons to suspect that human fact-finders would implement prohibited inference rules, there are corresponding reasons of evidential consideration not to rely on human fact-finders. We can even concede, for the sake of argument, that these reasons may even be of equal strength to the reasons that decision-makers have to avoid relying on black box systems (though see below). This doesn't show, though, that the reasons to avoid both approaches to decision-making cancel out, neutralizing our argument for the Explainability Thesis. There is a third option available—using interpretable predictive models—that avoids the problems we have identified to a significant extent.

As we mentioned above, decades of research have found that, across a wide variety of domains, even simple linear models often outperform human experts at predictive tasks, and interpretable models often perform about as well as black box models (Bell et al., 2022; Rudin, 2019). This suggests that interpretable predictive models will often be a viable alternative to both black box systems and human decision-makers in terms of overall performance. Moreover, for reasons that we have already seen, interpretable models are less likely to inadvertently implement prohibited inference rules than black box systems. On the one hand, they are less likely to implement a prohibited inference rule in the first place. The fact that they are trained using less flexible statistical learning methods and perform computations over fewer features of decision-subjects means that they are less likely to overfit their training data or exploit morally inadmissible evidence that is not explicitly encoded. And the fact that they are not as data-hungry as black box systems means that they are less likely to make inaccurate predictions due to data quality issues. On the other hand, in the event that they do end up implementing a prohibited inference rule, such as one that exploits morally inadmissible evidence or ignores readily available evidence that would benefit decision subjects, the problem will be easier for decision-makers to safeguard against, because it will be easier to detect.

We are happy to concede, then, that our arguments generalize to human decision-makers, and so that decision-makers will often have reasons of evidential consideration to avoid basing decisions on both black box systems and humans exercising their judgment. Our arguments do not generalize as strongly to interpretable systems, though, which suggests that using an interpretable system will often be the best way to show evidential consideration.

(2) While our arguments suggest that decision-makers often have reasons of evidential consideration to avoid relying on human decision-makers, those reasons are not necessarily as strong as their reasons to avoid relying on black box systems. First, different moral standards may apply to human- and machine-based decision systems in virtue of morally significant differences between the two types of

systems.[45] Second, if we allow even the most modest possibility that human decision-makers can evaluate what evidence they are responding to and how, then there will be a morally relevant asymmetry between relying on human decision-makers and relying on black box systems. Whether our arguments provide similarly strong reasons to eschew black box systems and human decision-makers thus remains an open question.

7 Duties of practical consideration

Whereas duties of evidential consideration constrain fact-finding, duties of practical consideration constrain decision-making—the task of deciding how to treat decision-subjects given the results of fact-finding. For example, the fact that an employer has promised to give a newly created position to a particular employee generates a reason for the employer to give that employee the role and a corresponding duty of practical consideration to give the promise appropriate weight during the hiring process. This is not a duty of evidential consideration, as it does not pertain to fact-finding regarding the subject's features.

So far, we have focused on the use of black box systems in fact-finding. However, a black box system can also be used in decision-making, implementing decision rules rather than inference rules.[46] Indeed, there is a growing interdisciplinary field—machine ethics—that aspires to build machines that can simulate the practical reasoning capacities of human agents by implementing suitable decision rules (Anderson and Anderson 2010). For example, Susan and Michael Anderson have experimented with using machine learning to infer decision rules underlying the moral reasoning of expert bioethicists about how clinicians ought to resolve moral dilemmas, and then programming caregiving robots to implement those rules (Anderson & Anderson, 2010). The Andersons' experiments used interpretable machine learning methods, but of course black box machine learning methods could be used instead, resulting in decision systems that are not rule transparent.[47]

[45] One difference between human- and algorithm-driven decision systems that is often emphasized, for example, is that once a decision-making algorithm has been developed, it can be deployed at scale, doing the work of countless human decision-makers. This might be taken to justify holding black box systems to different standards than human-driven systems (O'Neil 2016; Creel and Hellman 2022). See Zerilli et al. (2018) for objections to this view. We consider a quite different way in which automated and human decision-making systems may differ morally in Section 7.

[46] Note that an automated system might collapse fact-finding and decision-making into a single process rather than two discrete processes.

[47] It may not even be possible to build fully rule transparent systems that are capable of fully simulating the moral reasoning capacities of human agents. McDowell (1979) argues that it is "quite implausible that any reasonably adult moral outlook admits of … codification" in terms of a relatively compact set of explicit decision rules applied mechanically. Purves et al. (2015) argue that if this "anti-codifiability thesis" is correct, then AI-based systems will be unable to "adequately replicate" the moral reasoning capacities of human agents, as AI-based systems are only capable of making decisions by following "a discrete list of instructions provided by humans" (p. 857). This argument fails as applied to black box machine learning systems for two reasons: (1) the decision rules they apply are not hand-coded, but inferred from

We will consider two ways in which relying on a black box system in decision-making might lead to failures of practical consideration. First, black box systems that implement decision rules (as opposed to inference rules) are liable to implement decision rules that are not a morally acceptable basis for decision-making. Second, decision-makers are sometimes obligated to decide how to treat decision-subjects by exercising their capacities as full-blown moral agents, rather than outsourcing decision-making to a system that lacks these capacities.

7.1 Decision rules and duties of practical consideration

Like the inference rules discussed above, decision rules may be implemented by human decision-makers or automated systems. And just as some inference rules may be morally prohibited in virtue of decision-subjects' claims on how fact-finding should work, some decision rules may be morally prohibited in virtue of decision-subjects claims' on how decision-making should work. Continuing our earlier example, if our employer decided to use a black box system to decide which employee to hire for the role, but that system's decision rules did not treat the fact that one employee was promised the job as relevant, then that would count as a failure of practical consideration resulting from a failure to implement permissible decision rules.

We take it to be obvious that it will often be impracticable to fully anticipate in advance (a) what kinds of moral claims particular decision-subjects might have on how they ought to be treated and (b) how those moral claims might interact with claims others have that are relevant in context to determine what should be done. This, in conjunction with the fact that decision-makers cannot simply inspect the decision rules that a black box system is implementing, means that it will often be impracticable to design a black box system that decision-makers can be confident does not implement prohibited decision rules, just as it is often impracticable to ensure that black box systems will not implement prohibited inference rules. As a result, decision-makers will often have a duty of practical consideration not to base decision-making on the outputs of a black box system, because doing so would create a risk that they will fail to respond adequately to decision-subjects' moral claims on how decision-making is conducted.

To illustrate how decision-subjects' claims against being subjected to prohibited decision rules might give rise to a duty to avoid relying on black box systems, let us posit a constraint on decision-making based on the Kantian injunction against treating people as mere things.[48] One gloss of the Kantian injunction concerns how we

Footnote 47 (continued)

examples; and (2) they have the capacity to learn decision-making strategies that are too nuanced to spell out in the form of a compact set of explicit rules. However, this leaves open the possibility that *rule transparent* AI systems are inherently incapable of fully replicating the moral reasoning of adult humans.

[48] We think that our view about how duties of consideration interact with black box decision systems is consistent with a wide range of normative theories and other normative commitments. We've chosen to illustrate these interactions using the Kantian injunction for concreteness. It is also for this reason that we don't belabor a defense of the injunction or justify a move from the Strawsonian conception of the injunction to a version of the injunction as a constraint on decision rules.

explain the behavior of others; it is a requirement that when engaging with others "we must think of them as agents, not merely as causal or statistical objects" (Rini, 2020 p. 369).[49] Consider the relationship between this idea and recent scholarship developing Strawson's suggestion that we owe it to others to interpret their behavior by adopting the *participant stance* (Rini, 2020; Schroeder, 2019; Strawson, 1962). According to Strawson, we adopt the participant stance towards someone when we attempt to explain their behavior in terms of "reasons rather than causes"—that is, when we attempt to interpret their behavior as the product of their capacity to act rationally, as opposed to the product of arational causal influences. I might inappropriately treat a person as a thing by failing to adopt the participant stance toward her when I ought to. For example, I might credit to her parents all the responsibility for her flourishing and achievements, treating each action she undertakes in adulthood as *nothing other than* an event in a causal chain tracing back to her upbringing. This would treat her as a mere thing, rather than an agent autonomously contributing to her own life.

We can imagine a constraint on decision-making inspired by this Strawsonian conception of the Kantian injunction. Let us take the Kantian injunction to be a constraint on which descriptive properties of others we may rely on when making decisions about them. We treat decision-subjects as mere things, on this interpretation, when our decisions about them rely too heavily on features disconnected from their agency.[50]

What implications does the Kantian injunction, understood in this way, have for whether basing decisions on a black box system would be consistent with due consideration? That depends.

First, decision-makers may be required in some contexts to ensure not only that *they* reason about decision-subjects in a way that complies with the injunction, but also that any decision system they *rely* on complies with the injunction. It seems plausible, for example, that a military commander might violate the injunction by knowingly delegating decision-making about matters of life and death to someone who is incapable of understanding others as agents. Suppose we are in such a context, and are contemplating whether to rely on a particular black box system during the decision-making process. It is hard to see how we could be confident that the system's decision rules comply with the Kantian injunction. Even if we are confident that the system's input data does not *explicitly* represent features lacking an appropriate connection to decision-subjects' agency, such features might nonetheless be *implicitly* represented in a way that allows the system to infer and exploit them. And since black box systems are not rule transparent, we cannot rule out this possibility by simply inspecting the system's decision rules. We therefore have a pro

[49] What the actual injunction amounts to, and how we can satisfy the injunction while recognizing that people are, in a deep way, causal and statistical objects, is an open question (Schroeder 2019). Our analysis here should be consistent with a variety of possible interpretations.

[50] Another way to understand the Kantian injunction is as a constraint on inference rules, telling us, for example, what constitutes morally admissible evidence. (This is one way to interpret Eidelson 2013; see footnote 29 above.) We think it is more naturally understood as a constraint on decision rules, but encourage those that prefer the evidential interpretation to see our argument here as further developing the argument made in Sect. 5.3.

tanto reason of due consideration, grounded in the Kantian injunction, not to rely on the system. The more general lesson here is that relying on a black box system may interfere with decision-makers' ability to determine whether they are basing decisions on morally prohibited decision rules. (This point is analogous to points made above about black box systems inadvertently implementing prohibited inference rules.)

However, it is important to note that there are contexts in which decision-makers are *themselves* obligated to reason in a way that satisfies some constraint, but are nonetheless permitted to *delegate* decision-making to proxies that are not thus constrained. For example, even if we assume that legitimate use of the state's coercive power requires that its policies be justified by public or neutral reasons, the state may appeal to such reasons to justify policies giving more proximal decision-makers discretion to decide on the basis of non-public or non-neutral reasons. For example, the state may legitimately give discretion to the National Science Foundation to make decisions about which basic science to fund *even if* the NSF's reasons won't satisfy publicity or neutrality requirements—precisely because allowing such discretion yields public goods that serve to legitimate it (Brighouse, 1995).

There are important lessons to be drawn from this, but they do not undermine our arguments in this section. First, even in cases where it is permissible for a decision-maker bound by some constraint on decision-making to hand off decision-making to a proxy that is not so constrained, it does not follow that there are no *other* constraints on the decision rules the proxy may implement.[51] Second, the foregoing discussion suggests that different decision-makers or decision-making systems may be subject to different moral constraints in virtue of their differing capacities and their differing relationships to decision-subjects. That may sound nearly platitudinous, but it seems underappreciated by those that worry about holding black box systems to double standards.

7.2 Beyond decision rules: duties of agential consideration

In the rest of the paper, we will focus on a second way in which basing decision-making on a black box system can result in failures of practical consideration. The duties of consideration that we have discussed so far all pertain to what is sometimes called the "decision logic" of the decision-making system, which is jointly constituted by the inference and decision rules that it implements. These duties do not directly constrain what kind of system implements those rules, but only the content of the rules themselves. As a result, the duties of consideration (evidential and practical) that we have discussed could in principle be satisfied by relying on any sort of decision-making system—one where the rules are implemented by human decision-makers, an automated system, or some combination. The trouble, as we have argued, is that it is difficult in practice to design a black box system that can be trusted to implement appropriate rules.

[51] For example, we could imagine that there is a public reasons justification for relying on a decision-making proxy in a given context only if that proxy itself can be trusted to abide by the Kantian injunction.

By contrast, what we will call *duties of agential consideration* do place constraints on the nature of the system making the decisions. In cases where they apply, these duties require that decision-making be carried out by full-blown moral agents exercising their powers of moral reasoning, and that those agents deliberate in good faith to reach a decision that respects the decision-subject's moral claims on the decision-making process.

To motivate the existence of duties of agential consideration, consider the following thought experiment:

> Computer scientists announce that they have discovered a method to create customized models for any individual eligible to serve on a jury. The models are trained on personalized data sets and can predict with perfect accuracy how a given juror would find in any given criminal case by implementing the inference and decision rules of the modeled individual. After years of testing, the court system adopts the Juror Substitution Policy. The policy requires that individuals be called for jury duty using the usual method: They come to court, lawyers are given a chance to evaluate and dismiss them, and so on. However, once jurors are selected, they may leave and their juror model will be used to adjudicate the case. Imagine that these models take as inputs whatever written, visual, or auditory information a human juror would process during a criminal trial, and perfectly replicate the judgments their human counterparts would make in light of such information (including instructions from the judge to disregard certain information). Further, imagine that these models reach their judgments by implementing the same inference and decision rules that their human counterparts would have used.

We submit that there is something morally problematic about the use of juror models. However, the wrong cannot be explained in terms of the nature of the inference or decision rules that the trial system implements. By hypothesis, juror models implement the same inference and decision rules that their human counterparts would have. Nor can the wrongness be explained by appeal to duties of transparency. Jurors and juror models offer up the same sort of information to decision-subjects: a verdict. Furthermore, the use of such models strikes us as objectionable even if decision-subjects had access to a trove of information regarding the "deliberations" of the models, satisfying whichever duties of transparency one might prefer.

What, then, is the problem? In our view, at least part of what is morally problematic with the use of juror models is that certain important decisions—such as decisions about whether to impose criminal punishment—normally ought to be made by full-blown moral agents exercising their distinctive moral capacities with a level of care that is appropriate given the stakes.[52] When a human decision-maker makes a decision about how to treat a decision-subject by carefully reasoning through what claims the decision-subject has and how those claims bear on how they ought

[52] Another problem is that, at least in some countries, defendants are legally entitled a trial by a jury of their peers, which is to say their fellow citizens. Insofar as this legal entitlement has moral force, it also partly explains why relying on juror models would be problematic. The general point that we are after in the main text, though, does not depend on a connection to legal entitlements.

to be treated, she thereby takes on a special kind of responsibility for the resulting decision—one that she would not have had she delegated decision-making to another person or automated system. Further, by owning the decision in this way, she thereby demonstrates an important kind of respect for the decision-subject: she both recognizes and gives appropriate weight to their status as a fellow member of her moral and political community in her deliberations.[53]

This helps to explain why the Juror Substitution Policy seems problematic: it replaces decision-makers that are fellow members of the defendant's moral and political community and who are capable of exercising agential consideration towards the defendant with automated systems that are not and cannot. When human jurors decide whether a defendant ought to be convicted and punished, they take responsibility for the defendant's punishment (and designation as a criminal) on behalf of the broader polity, and thereby demonstrate the polity's respect for the defendant's status as a fellow citizen. This, in turn, helps to legitimate the defendant's change in criminal status and ensuing punishment (in the case of conviction). When juror models are used, by contrast, there is no member of the polity that intentionally takes on this kind of direct responsibility for the decision. This demonstrates a morally objectionable lack of respect for the defendant's moral and civic status.

We suspect that a similar argument can help diagnose concerns about responsibility gaps that arise in the context of autonomous systems (Asaro, 2020; Matthias, 2004; Roff, 2013; Sparrow, 2007).[54] As an example, consider Sparrow's (2007) well-known argument that deploying lethal autonomous weapons (LAW) with sophisticated decision-making capacities is impermissible because it would lead to "responsibility gaps": situations in which someone *ought* to be held responsible for a LAW killing an illegitimate target, but no suitable candidates exist (because the LAW itself is not a moral agent and no moral agent had suitable control over the LAW's actions). This argument is vulnerable to the rebuttal that—as Sparrow himself recognizes—accidental civilian casualties that no one is directly responsible for are inevitable in war. Why would accidental deaths resulting from the decisions of an elaborate piece of software be any worse than accidental deaths that arise from other causes, such as bad intelligence or equipment malfunctions?

The answer, we suggest, is as follows. The decision to take someone's life is the kind of decision that we are normally obligated to make only after exercising agential consideration as carefully as circumstances allow.[55] Delegating such decisions to a piece of software that is incapable of agential consideration fails to provide potential victims with the agential consideration that they are owed, and so seemingly fails to show them the respect they deserve as members of the moral community. So, the difference between a LAW deciding to kill illegitimate targets and other kinds

[53] The ideas we develop here regarding agential consideration are, we think, closely related to some of the ideas presented in (Rubel et al., 2021).

[54] For a discussion of the relationship between responsibility gaps and XAI, see (Baum et al., 2022). For a more skeptical take on the problem of responsibility gaps see (Hindriks and Veluwenkamp 2023; Tigard 2021).

[55] This provides a substantive way to fill out a suggestion by Purves et al. (2015) that making certain kinds of decisions, like those made by a soldier about whether to kill, requires the exercise of moral judgment.

of accidental casualties in war is that decision-making authority has been delegated to the LAW. When a bomb malfunctions and hits a civilian target, this is not the product of a similar delegation of decision-making authority. The problem is not so much that no one is responsible for the deaths as that responsibility for deciding whether to kill was inappropriately delegated.[56]

Whatever one thinks about the case of LAWs, we take it that duties of agential consideration are part and parcel of what it means to be appropriately responsive to the distinctive moral status of persons in a wide range of contexts. What it means to be "appropriately responsive" to a particular entity's moral status in a particular context depends on various details about the capacities of that entity as well as our relationships to it (Sandler & Basl, 2021).[57] However, given the capacities persons typically have and the kinds of relationships we typically have to one another, we often owe each other agential consideration.

Consider having to make a decision on behalf of your partner about something consequential, such as how to manage a medical emergency while they are unconscious or a decision about whether to accept a time-sensitive offer while they are on a long flight. Consider also political representatives tasked with making trade-offs between various interests of their constituents, or financial advisers making decisions about the stock portfolios of unsophisticated or inattentive clients. In each of these contexts, we plausibly owe agential consideration to others, though exactly what agential consideration requires differs from context to context. In the case of juries, jurors' duties of agential consideration are mediated by the law; jurors are to exercise their agential capacity in their role specifically as jurors and not as unrestricted moral agents.[58] By contrast, financial advisers' duties of agential consideration to their clients may be mediated by fiduciary duties, laws applicable to financial institutions, etc. And our duties of agential consideration to our partners are mediated by the details of our shared histories and the specific nature of our relationship to them. What is constant across these cases is that a failure to exercise our agential capacities appropriately is a failure to be appropriately responsive to the moral status of the relevant decision-subjects.[59]

[56] We do not here take a stance on when it is appropriate or inappropriate to delegate these kinds of decisions to those that lack agential capacities.

[57] For example, what it means to be appropriately responsive to the moral status of a pet dog and a wild coyote differs greatly, despite their similar capacities (Palmer 2010).

[58] Indeed, it is in virtue of this mediation by the law and political institutions that the exercise of agential capacity by jurors plays the additional role of legitimating the verdicts of jury trials. However, we think it would be a mistake to think that jurors' obligations were solely a function of the law. Consider, for example, the phenomena of jury nullification, whereby jurors exercise their sense of justice to acquit a defendant despite the defendant having violated some (presumably unjust) law. Apparently, then, what constitutes due consideration by jurors is not solely a function of the legal apparatus. We thank an anonymous referee for this journal for pushing us to acknowledge and grapple with the relationship between the prescribed legal responsibilities of jurors and what we are describing as their duties of agential consideration.

[59] Though see the following section for an important caveat about the strength of this claim.

7.3 Agential consideration and the explainability thesis

We are now in a position to explain how relying on a black box system might interfere, in various ways, with the duty to show agential consideration.

Outsourcing decision-making wholesale to such a system is incompatible with showing agential consideration to decision-subjects for the simple reason that black box systems *cannot* show agential consideration: only full-blown moral agents can do that, and automated systems are not full-blown moral agents.[60] Substituting a black box system for human decision-makers is therefore at least *prima facie* impermissible in cases where decision-subjects are owed agential consideration, such as in jury trials.

Notably, the reasons grounded in duties of agential consideration that tell against ceding decision-making authority to black box systems also tell against ceding such power to *any* automated system. In cases where agential consideration is owed, the distinction between black box systems and automated systems based on simpler predictive models is largely irrelevant. What about "human-in-the-loop" (HITL) decision-making structures—those involving predictions or recommendations issued by black box systems that are fed to a human with final authority (Bell et al. 2020)? It is easy to see that the mere inclusion of a human is not sufficient to ensure agential consideration. If the human defers to the black box system's recommendation without further thought, then there is no meaningful difference between a decision structure that includes the human and one that does not. At the other extreme, there is little reason to doubt that a human *could* give full agential consideration after consulting the recommendation of a black box system. A judge who takes the time to carefully examine the details of a defendant's circumstances is not rendered *incapable* of showing agential consideration simply by consulting a black box system's recommendation.

What more can be said about HITL structures, beyond these observations about extreme cases? On this question, we must largely demur. As we have seen, agential consideration may be owed across a wide range of contexts and for widely varying reasons. That complexity will presumably give rise to some variability with respect to what discharging particular duties of agential consideration requires. But because we know that blind deference to an automated system is inconsistent with agential consideration, we can at least conclude that HITL decision-making structures introduce some risk that humans will fail to give agential consideration in contexts where

[60] Note that on particularly reductionist pictures of agency, deliberation, etc., like those favored by some participants in debates about algorithmic transparency, black box systems could themselves meet the requirements of agential consideration (Zerilli et al. 2018 and 2022). However, even if we assume that black box systems are moral agents in a limited sense, two problems remain for defending the idea that they can satisfy duties of agential consideration. First, the reductionist view of agency provides no special reason for believing that automated systems are capable of *considering decision-subjects as people* and deciding how to treat them by *considering the moral implications of that status*. These are sophisticated cognitive achievements; attributing them to present-day black box systems would be a wild overinterpretation of what's happening. Second, decision-subjects are plausibly owed agential consideration from *certain kinds of full-blown moral agents*—such as fellow citizens, a representative of the company, etc.—and not just *any old moral agent*. For example, criminal defendants are plausibly owed the agential consideration of their fellow citizens, rather than citizens of a different country or sentient machines.

it is required.[61] To the extent that our evidence suggests that the risk of inappropriate deference is heightened when black box systems in particular are used, our duties of agential consideration may provide special reason to resist the use of HITL structures incorporating black box systems.

Finally, let us return again to the point that different decision-making systems might be under substantially different normative constraints, grounding asymmetries in the transparency demands we should make of them. We do not take duties of agential consideration to provide a decisive reason against deploying algorithmic decision-making systems, even black box ones. For example, there may be scenarios in which the advantages offered by juror substitution outweigh attendant failures to meet duties of agential consideration. However, notice that we can justify different requirements of transparency for jurors and for juror models. It might be reasonable to allow human jurors to deliberate in secret: even if we have strong reasons to require transparency (e.g., because it would help prevent juror misconduct), those reasons might be outweighed by even stronger reasons against transparency (e.g., because it would render jurors vulnerable to manipulation). This justification for secrecy, though, would not apply to juror models. Just as with decision rules, attending to the situatedness of decision-makers and the different ways that moral considerations apply to them helps us see that differential transparency requirements need not constitute an objectionable double standard.

8 Conclusion

Our duties to decision-subjects—including our duties to implement permissible inference and decision rules, and our duties to provide agential consideration—often give us significant reasons to reject decision-making systems based on black box AI systems. Sometimes this is because we can't verify whether such systems abide by these duties, other times it is because they can't possibly do so, and other times it is because integrating them into decision systems undermines our ability to do so. These duties not only ground the Explainability Thesis, but also help us to see what forms of transparency would serve to help us realize our duties to decision-subjects in particular contexts and why there are often good reasons to hold human decision-makers and automated decision systems to different standards.

Unfortunately for those seeking to defend broad transparency standards or sweeping claims about the impermissibility of using black box systems, recognizing the spectrum of moral duties that ground the Explainability Thesis reinforces the lesson that the import of our design decisions regarding automated decision systems is highly context-sensitive. However, we also think that these arguments provide motivation for further philosophical work. For example, there is likely much to be learned from thinking about the decisions we make in our interpersonal relationships and the constraints on those decisions, and it is essential to think more carefully

[61] The empirical literature on "automation bias" suggests that this risk is significant (Citron 2008).

about the ethics of delegating decision-making to others who are not bound by the same constraints.[62]

Funding Open access funding provided by Northeastern University Library. A portion of John Basl's efforts on this project were funded by NSF Award 1928654.

References

Anderson, M., & Anderson, S. (2010). Robot be good: A call for ethical autonomous machines. *Scientific American*. https://www.scientificamerican.com/article/robot-be-good/.

Anderson, M., & Anderson, S. (Eds.). (2011). Machine Ethics. Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.

Asaro, P. (2020) Autonomous weapons and the ethics of artificial intelligence. *Ethics of Artificial Intelligence, 212*.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *California Law Review, 104*, 671–732.

Basl, J., & Sandler, R. (2021) *Getting from commitment to content In AI and Data Ethics: Justice and Explainability.* Steven Tiell, Managing Editor. *Atlantic Council*. https://www.atlanticcouncil.org/in-depth-research-reports/report/specifying-normative-content/

Baum, K., et al. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology, 35*(1), 12.

Beeghly, E. (2018). Failing to treat persons as individuals. *Ergo, 5*(26), 687–711.

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 248–266).

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013.

Bolinger, R. J. (2021). Explaining the justificatory asymmetry between statistical and individualized evidence. In: *The social epistemology of legal trials* (pp. 60–76). Routledge.

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *BMJ, 333*(7562), 284.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.

[62] We were fortunate to receive helpful feedback from many people on earlier versions of this paper. For close attention to the details of previous drafts, special thanks is owed to Jerome Hodges, Milo Phillips-Brown, Schuyler Sturm, and two anonymous referees for this journal. We also received very useful input from Beba Cribalic, Will Fleischer, Lily Hu, Gregory Keenan, Vance Ricks, Nikita Shepard, and Matthew Smith along the way. Finally, we are grateful to audiences at the following venues: the Jain Family Institute, the Digital Life Institute Seminar Series (Cornell Tech), the University of Georgia Department of Philosophy, the 13th Annual Rocky Mountain Ethics Congress (University of Colorado Boulder), the Ethics and Technology Seminar Series (University of Macerata and University of Rome 3), the Law and New Technologies Conference (University of Catanzaro), the Hoffman Center of Business Ethics (Bentley University), and the University of Wisconsin-Madison Department of Philosophy.

Brighouse, H. (1995). Neutrality, publicity, and state funding of the arts. *Philosophy & Public Affairs, 24*(1), 35–63.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*(12), 5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass, 14*(10), e12625.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence, 2*, 731–736.

Buckner, C., & Garson, J. (2019). Connectionism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019a Edition). https://plato.stanford.edu/archives/fall2019a/entries/connectionism/.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512.

Caruana, R., et al. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.*

Castro, C. (2019b). What's wrong with machine bias. *Ergo, an Open Access Journal of Philosophy, 6*, 1.

Citron, D. K. (2008). Technological Due Process. *Wash. UL Rev., 85*, 1249.

Clinciu, M., & Hastie, H. (2019). A survey of explainable AI terminology. In *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence* (Tokyo, Japan) (NL4XAI 2019), Jose M. Alonso and Alejandro Catala (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 8–13. https://doi.org/10.18653/v1/W19- 8403.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. arXiv preprint arXiv:1808.00023.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568–589.

Creel, K., & Hellman, D. (2022). The algorithmic Leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy, 52*(1), 26–43.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *The Washington Post*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Di Bello, M., & O'neil, C. (2020). Profile evidence, fairness, and the risks of mistaken convictions. *Ethics, 130*(2), 147–178.

Dutta, S., Wei, D., Yueksel, H., Chen, P. Y., Liu, S., & Varshney, K. (2020). Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In International Conference on Machine Learning (pp. 2803–2813). PMLR.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

Eidelson, B. (2013). Treating people as individuals. In Deborah, H., Sophia, M. (Eds.) *Philosophical Foundations of Discrimination Law*. Oxford University Press.

Eidelson, B. (2015). *Discrimination and disrespect*. Oxford University Press.

Equivant, Inc. (2019) Practitioner's Guide to COMPAS Core. https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf.

Enoch, D. (2016). II—What's wrong with paternalism: Autonomy, belief, and action. *Proceedings of the Aristotelian Society, 116*(1), 21–48.

Enoch, D. (2018). In defense of procedural rights (or anyway, procedural duties): A response to Wellman. *Legal Theory, 24*(1), 40–49.

Enoch, D., & Spectre, L. (2021). Statistical resentment, or: What's wrong with acting, blaming, and believing on the basis of statistics alone. *Synthese, 199*(3), 5687–5718.

Feinberg, J. (1974). Noncomparative justice. *The Philosophical Review, 83*(3), 297–338.

Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme, 19*(4), 18.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.

Grant, D. G. (2023). Equalized odds is a requirement of algorithmic fairness. *Synthese, 201*(3), 1–25.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*.

Hellman (2023). Big Data and Compounding Injustice. *Journal of Moral Philosophy*.

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese, 201*, 21.

Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences, 113*(16), 4296–4301.

Howell, J., & Korver-Glenn, E. (2018). Neighborhoods, race, and the twenty-first-century housing appraisal industry. *Sociology of Race and Ethnicity, 4*(4), 473–490.

Hu, L. (forthcoming). What is "race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R (second edition)*. Springer.

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese, 198*(10), 9941–9961.

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2020). Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (statistics in Society), 183*(3), 771–800.

Kim, P. T. (2016). Data-driven discrimination at work. *Wm. & Mary L. Rev., 58*, 857.

Krishnan, M. (2019). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology, 33*(3), 487–502.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473.

Lippert-Rasmussen, K. (2011). "We are all different": Statistical discrimination and the right to be treated as an individual. *The Journal of Ethics, 15*(1–2), 47–59.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*, 175–183.

Mayson, S. G. (2018). Dangerous Defendants. *Yale LJ, 127*, 490.

Mayson, S. G. (2019). Bias in, bias out. *The Yale Law Journal, 128*(8), 2218–2300.

McDowell, J. (1979). Virtue and reason. *The Monist, 62*(3), 331–350.

McKinney, et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*, 89–94.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review, 55*, 3503–3568.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Palmer, C. (2010). *Animal ethics in context*. Columbia University Press.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Rawls, J. (1999). *A theory of justice: Revised edition*. Harvard University Press.

Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence, 3*(10), 896–904.

Roff, H. M. (2013) Killing in war: Responsibility, liability, and lethal autonomous robots. Routledge Handbook of Ethics and War (pp. 352–364). Routledge.

Rini, R. (2020). Contingency inattention: Against causal debunking in ethics. *Philosophical Studies, 177*, 369–389.

Rubel, A., Casto, C., & Pham, A. (2021). *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
Sandler, R., & Basl, J. (2021). Justified Species Partiality. In Bovenkerk, Keulartz (Eds.) *Animals in our Midst*, The International Library of Environmental, Agricultural and Food Ethics, 33.
Scanlon, T. (1975). Thomson on privacy. Philosophy & Public Affairs, 315–322.
Scanlon, T. (2018). *Why does inequality matter?* Oxford University Press.
Schroeder, M. (2019). Persons as things. In M. Timmons (Ed.), *Oxford Studies In Normative Ethics* (Vol. 9). Oxford University Press.
Schwitzgebel, E. (2019). Introspection. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). https://plato.stanford.edu/archives/win2019/entries/introspection/.
Selbst, A., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*(3), 1085–1139.
Shellenbarger, S. (2019). Make Your Job Application Robot-Proof. *The Wall Street Journal*. https://www.wsj.com/articles/make-your-job-application-robot-proof-11576492201.
Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.
Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2239–2250).
Strawson, P., (1962). Freedom and Resentment. *Proceedings of the British Academy* 48: I-52.
Thomson, J. J. (1986). Liability and individualized evidence. In W. Parent (Ed.), *Rights, restitution, and risk* (pp. 225–250). Harvard University Press.
Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology, 34*(3), 589–607.
Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy, 30*(2), 209–229.
Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law. *West Virginia Law Review, 123*(3), 735–790.
Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683.
Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science, 89*(1), 1–19.

Authors and Affiliations

David Gray Grant[1,2] · Jeff Behrends[3] · John Basl[4]

✉ John Basl
j.basl@northeastern.edu

David Gray Grant
david.grant@ufl.edu

Jeff Behrends
jbehrends@fas.harvard.edu

[1] University of Florida, Gainesville, USA

[2] Jain Family Institute, New York, USA

[3] Harvard University, Cambridge, USA

[4] Northeastern University, Boston, USA