Ethics and Information Technology (2026) 28:18
https://doi.org/10.1007/s10676-026-09892-3

ORIGINAL PAPER

All you need is…. justification: algorithmic justifiability trumps transparency

Anantharaman Muralidharan[1] · **Julian Savulescu**[2,3]

Abstract
Most ethical guidelines on AI tout algorithmic transparency, the openness of an algorithm's inner workings to human scrutiny, as an important desideratum in algorithmic deployment. Algorithmic transparency has been touted as important for valuable goals like procedural fairness, AI trustworthiness, contestability and planning around AI decision-making. This paper argues that these goals are better served by a distinct desideratum, algorithmic justifiability, the ability of an algorithm to provide understanding about why the algorithm's decision is correct.

Keywords AI · Transparency · Justifiability · Procedural Fairness · Trust · Contestability

Introduction

Most, if not all accounts of the ethical deployment of algorithms, found in institutional, national and international guidelines and policies (Hagendorff, 2020), posit that algorithms need to be transparent (i.e. either. Some exemplars include the EU AI Act (European Parliament 2024), Beijing AI principles (Beijing Academy of Artificial Intelligence, 2019), the US's Report on the Future of Artificial Intelligence (Holdren et al., 2016) and the Singapore Bioethics Advisory Committee's consultation paper (Tan et al., 2023). Algorithmic transparency is often touted as required or conducive to various ends like ensuring that decisions using AI are procedurally fair (that is, relevantly similar cases are treated alike), trustworthiness of AI, enabling AI assisted decisions to be challenged and planning future courses of action in order to avoid adverse decisions (Babic & Cohen, 2023; Selbst & Barocas, 2018). While transparency has been invoked by some (Loi et al., 2021; Watson & Floridi, 2021; Zednik & Boelsen, 2022) as conducive towards justifying algorithmic decision-making, recent work has suggested that they are distinct desiderata (Muralidharan et al., 2024). Whereas transparency is concerned with making explicit how the algorithm reached a given decision, justifiability is concerned with helping the users and other related parties understand why the decision is correct (if it is indeed correct). Even when decisions are correct, the algorithm may not have arrived at the decisions on the basis of those considerations that make that decision correct. That is to say, the questions of how the algorithm reached a decision and what makes the decision correct can often come apart. This means that transparency, which merely reveals how an algorithm reached its result does not tell us whether the result in question is correct or help us understand why this would be the case. In this paper, we go further and argue that in addition to being a distinct desideratum, all the reasons we might have in favour of algorithmic transparency are in fact reasons for algorithmic justifiability instead. That is to say, algorithmic justifiability does at least as good a job at achieving the goals people aim at when they pursue transparency. As such, an algorithm being justifiable obviates any reason for it to be transparent. Therefore, we should pursue algorithmic justifiability instead of transparency.

Our strategy in this paper is as follows: In Algorithmic justifiability vs. transparency, we will rehearse the distinction between algorithmic justifiability and transparency with the aid of a hypothetical AI model called Justifiable AI. In Procedural justice, we show why justifiable AI can deliver

✉ Anantharaman Muralidharan
murali@nus.edu.sg

✉ Julian Savulescu
julian.savulescu@uehiro.ox.ac.uk

[1] National University of Singapore, Singapore, Singapore

[2] University of Oxford, Oxford, UK

[3] National University of Singapore, Singapore, Singapore

Published online: 23 March 2026

procedurally fair decisions at least as well as transparent algorithms. Planning to avoid bad outcomes explores how justifiable AI better helps people affected by the decision avoid bad outcomes than transparent AI. In Challenging decisions, we show that if we want challengeable decisions, pursuing transparent AI would be a mistake whereas justifiable AI would be exactly what is required. In Trust, we argue that algorithmic justifiability gives us more reason to rely on or trust an AI than algorithmic transparency. Conclusion examines what follows from the fact that justifiability trumps transparency.

Algorithmic justifiability vs. transparency

Before we proceed to the meat of the argument, let us clarify our terms. By algorithmic decision-making, we refer to human-in-the loop processes where an algorithm makes recommendations which a human may accept or reject. Algorithmic transparency refers to the openness of an algorithm to the scrutiny of its inner workings (Durán, 2021). At one extreme, white-box or interpretable algorithms like simple linear regressions and decision trees are those whose inner workings are completely open to scrutiny. It is in principle possible to see how the addition of any individual training datapoint will affect how the algorithm transforms any given input. A more limited form of transparency is exemplified in explainable AI. Explainable AI involves a primary black-box AI coupled with a post-hoc high-fidelity (Muralidharan et al., 2024) explainer algorithm, like LIME (Ribeiro et al., 2016), which, if successful[1], tells us some information about how the primary AI reached a given decision. This transparency is more limited because only information about how that decision was reached is revealed. This does not necessarily translate into how other, seemingly similar situations will be decided nor does it provide information about how the algorithm will evolve with the addition of new datapoints. Insofar as the post-hoc method genuinely reveals information about how a given output was generated, the algorithm is thereby partly transparent. We will, in this paper, regard post-hoc methods which give false or misleading information about how an output was generated as not explainable. The least transparent are black-box algorithms like random forests and neural networks which "resist knowledge and understanding" (Beisbart, 2021, p. 11643) of their inner workings.

Justifiability, by contrast, involves providing understanding whether a decision is correct (Elgin, 2017; Hills, 2009). A decision is correct whenever it satisfies the context appropriate normative criteria[2] (Muralidharan et al., 2024). For instance, in medical decisions, a treatment decision is correct only if the projected outcome of the decision promotes the patient's best interests and/or fits the patient's priorities and preferences to a sufficient degree. Another possible desideratum for correctness in clinical decisions may be whether the evidence available to the physician supports the claim that a given treatment option will have a certain projected outcome. And yet another desideratum is whether the distribution of benefits and burdens amongst a population of patients is fair. To generalise beyond the medical context, let us call these desiderata, that, when satisfied, make the decision correct, the correct-making features of the decision.

Notably, justifiability, as we will use the term here is not merely about epistemic justification[3], but more analogous to interpersonal justification. Interpersonal justification is richer than epistemic justification. Firstly, interpersonal justification is often thought to require an accessibility requirement[4] (Gaus 2011; Muralidharan 2023; Vallier 2011b): Propositions which are justifiable to a person must be accessible for them. In addition, across a range of cases, our actual practice of interpersonal justification requires more than just presenting sufficient evidence for believing a proposition. Consider that in a parole hearing, it is not enough to show that a particular psychiatrist is extremely reliable at predicting recidivism. If denied parole the prisoner and her loved ones would want an account of how it could be true that the prisoner is still a danger to society. Conversely, if released on parole, the public might desire to know how it could be true that the prisoner is no longer a danger to society.

[1] Defenders of interpretability like Babic (2021) and Rudin (2019) object that many so-called post-hoc methods do not actually reveal how the decision was reached. That just means that many post-hoc explainers are unsuccessful on their own terms. The argument in this paper is that even if successful, algorithmic transparency is not valuable for the reasons given in favour of it so far.

[2] These criteria, depending on the context, need not be moral considerations. The correct-making features of a decision about how to stack containers in a shipyard may be entirely non-moral like facts about schedules, container weights and other physical and logistical constraints. They are normative criteria in the sense that these facts make certain arrangements more efficient than others and it is desirable that shipyards operate efficiently.

[3] Contrast our view with Loi and colleagues (2021), according to whom certain kinds of transparency are required in order to be epistemically justified in believing that the decision is correct. Rather explicitly, they do not "require that individuals that are accountable for algorithmic decisions provide fully persuasive and non-corrigible justification" (Loi et al., 2021, p. 261). By contrast, we do require the provision of discursive reasons. Contrast, our view, also with Duran's according to whom all that is required for epistemic justifiability of the output is that the algorithm be sufficiently reliable (Durán forthcoming).

[4] We remain agnostic about a different accessibility requirement according to which the reasons underlying political decisions ought to be accessible to all reasonable persons (Vallier 2011a). It is possible that the same proposition might be justifiable to different people for different reasons which are not mutually intelligible.

Interpersonal justification of some proposition, at least in many contexts, thus seems to also require the provision of reasons (Baum et al., 2022) of a special kind: the truth-makers of the proposition. Where the proposition takes the form "this decision is correct", the truth-makers of that proposition just are the correct-making features of the decision.

While it may be that a treatment decision is correct only if it satisfies various desiderata, understanding, to a sufficient degree, why this is the case need not involve appreciating why all these desiderata obtain. For patients, for instance, understanding why a given decision is right for them is primarily a matter of seeing why the recommended treatment option fits their values and priorities (Muralidharan et al., 2024). Patients are often willing to trust that the physician's projections about the expected effects of the treatment fit the available evidence. By contrast, the physician would need to be able to understand why the projected outcome of a particular treatment is probable on the evidence. This evidence could include current medical and scientific theories, established medical practice and personal experience. Of course, if the physician were to present their reasons to the patient, this would increase the patient's understanding of why the treatment decision was correct. That said, the patient need not achieve the same level of understanding with regard to the empirical effects of the treatment decision as the physician in order to adequately, for her purposes, understand why the treatment decision is correct.

Put differently, a person understands why a decision is correct (assuming that it is) to the extent that her belief that it is correct is based on a grasp of the correct-making features of that decision. The physician understands why a given treatment might not extend the patient's life because she knows the patient's condition and is able connect the proposed treatment and patient's condition to the projected outcome reasoning using her medical knowledge. Meanwhile, the patient could be just as justified in believing that her life would not be extended because she obtains testimony from the physician. The testimony, in this case, does not confer understanding. The degree and aspects of understanding required may vary from context to context[5].

In order to illustrate the distinction between justifiability and transparency, consider the following hypothetical model we call Justifiable AI. Justifiable AI is a post-hoc model consisting of a primary, black-box AI, which makes recommendations, and a secondary model trained on relevant subject matter, including moral philosophy, as well as any other relevant source[6], which generates some argument or narrative which attempts to tell us why the decision by the primary output is correct (Muralidharan et al., 2024). It is a post-hoc model because the argument or narrative is generated on the basis of the output of the primary decision-making algorithm. Incidentally, it is also post-hoc in the sense that it attempts to give a post-hoc rationalisation of the primary AI's decision (Babic et al., 2021). Crucially, the post-hoc rationalisation may not necessarily correspond to how the algorithm actually reached the decision[7] (Turpin et al., 2023).

One of the key distinctions we wish to make is between explanatory or motivating reasons, on the one hand and normative or justifying reasons on the other (Smith, 1994). Motivating reasons specify the cause of an action or decision and hence explain that action or decision in causal terms. Normative reasons, on the other hand, specify what makes a decision or action correct and hence justify that action or decision. Importantly, motivating reasons need not be normative reasons. A terrorist might be motivated to bomb a synagogue because he hates jews but bombing the synagogue is certainly justified. Hatred of jews might be the terrorist's motivating reason, but cannot be anyone's normative reasons. Cashing out the distinction between Justifiable AI and a nearby model, Reason-Giving Explainable AI (RGXAI) (Baum et al., 2022), will also clarify what we mean by justifiability. According to Baum and colleagues (2022), RGXAI is a post-hoc model in which the secondary AI tries to generate the "motivating reasons" for the primary AI's decision. They note that these are not genuinely motivating reasons as AIs cannot have motivations as such. These ersatz motivating reasons are what would be motivating reasons if a person had made the decision for the cited considerations. Baum and colleagues see the provision of these "motivating reasons" as being able to provide insight into whether the decision accords with the normative reasons in favour of the decision. By contrast, the secondary algorithm in Justifiable AI, if it works well, directly provides the normative reasons in favour of the decision.

One might object that such post-hoc rationalisations do not have any justificatory force. The justificatory force comes entirely from knowing that an algorithm is reliable (Durán forthcoming; Loi et al., 2021). The post-hoc rationalisations seem superfluous, at best icing on the cake. To illustrate this point, one might compare two algorithms A and B: The first, A, is 99% reliable while the second, B, is 80% reliable. However, B also has a secondary algorithm

[5] We might think that trusting someone else allows us to distribute responsibility for understanding why the joint decision is correct (Muralidharan et al. forthcoming).

[6] This may go beyond philosophy textbooks because current attempts at developing approaches that integrate socio-ethical and structural aspects into AI ethics which may well be relevant to include in the training data for justifiable AI might not be part of philosophy textbooks in the nearest future.

[7] Arguably Chain-of-thought language models are also justifiable in this sense. They produce post-hoc rationalisations for their final answer.

which provides some narrative about why its output is correct. The key claim, according to the objection, is that despite B's secondary AI, if A and B each give different answers to a given question, we should defer to A's answer and reject B's answer.

However, things are not so clear cut. Depending on the content of the narrative, the right response to this disagreement may change. If the narrative presents genuine reasons that *demonstrate* that A got it wrong in this particular case, then we should defer to B instead. Even if A's decision is right 99% of the time in relevantly similar situations, it may still be that there is some discernible defeater for that decision. Indeed, it will be wrong in 1% of cases, and this might fall in that set. The likelihood that A is correct given the presence of the defeater can be significantly lower than 0.99. Depending on the nature of that defeater, it might even reduce to 0. If, for instance, there is conclusive or decisive evidence in favour of B's option, then A, despite its high antecedent reliability is likely mistaken in *this* particular instance. If this is right, then the post-hoc rationalisations can have justificatory force, provided that they are not bogus and identify genuine reasons as identified by a human being. In the 20% of cases where B gives the wrong result, the purported justification generated by the algorithm would be a weak or even bogus justification. It might involve some hallucination, fail to mention some available defeater or even involve non-sequiturs or other errors of reasoning. Detecting these errors should help users know when to reject the output of the primary algorithm.

At this point, some might object that we seem to be assuming that users will have perfect reasoning skills. However, this assumption is false; people do not have perfect knowledge and reasoning skills and may be convinced by fallacious but seemingly plausible narratives and rationalisations. Since some such narrative is generated by the justifiable AI regardless of whether the AI's primary output is correct, the rationalisation may even be harmful if it convinces people to defer to the AI when they should not.

Our response here is threefold. Firstly, while we acknowledge this risk, the existence of this risk is orthogonal to the point of whether the post-hoc rationalisation has justificatory force. The rationalisation has justificatory force insofar as it presents a genuine, non-misleading justification for the primary AI's output. Our ability to detect when it does this and distinguish such cases from cases where the purported justification is bogus may be very imperfect, but that is a separate question. Secondly, as a practical matter, algorithmic justifiability is not a substitute for algorithmic reliability. We can, in principle, have algorithms that are reliable *and* justifiable. We need not trade-off one against the other – typically justifiability will refer to statistical probabilities and reliability, but importantly, it can embrace reasoning that goes beyond this. Maximising reliability can reduce the chances of generating a bogus justification and, hence, reduce the chances of users being misled by such rationalisations. Thirdly, no matter how bad we are at detecting errors in reasoning, we are strictly worse at generating *de-novo* sound justifications than detecting errors in others' efforts to generate justifications. The nature of epistemic blind spots is that we are always especially bad at identifying our own blind spots. That is why we have peer review. Justifiable AI serves the same role, by potentially pointing out to users mistakes they have made in their reasoning and also giving them an opportunity to detect errors in the AI's output.

If the Justifiable AI is well trained, the physician and patient can follow the inferential connections presented by the secondary AI and evaluate whether the decision is indeed correct under the circumstances. Typically the secondary AI would base its argument on evidence of performance in an ecologically relevant setting, established theoretical knowledge, and relevant normative considerations such as patient values or theory of justice.

By contrast, an explainable AI is likely to present very different information. Rather than the correct-making features themselves, machine learning algorithms often use proxies that are correlated with the correct-making features to arrive at decisions. Algorithmic transparency thus can only deliver information about which proxy variables drove the decision. It does not necessarily tell us what these variables are proxies for. This is also why, contrary to claims made by some (Baum et al., 2022; Schmidt et al., 2025)[8], rather than algorithmic explainability, algorithmic justifiability is required for knowledge. Consider, for instance, the case described by Obermeyer and colleagues (2019) wherein an algorithm was used to allocate resources to patients in a healthcare system. Obermeyer found that the algorithm used expected medical expenditure as a proxy for medical need. Since the algorithm was fully transparent, they learned, for instance, how expected medical expenditure was calculated by the algorithm. They would not have thereby learned whether the algorithm distributed according to medical need, or whether the allocation decisions of the algorithm were, all things considered, correct.

However, unless one knows that the proxy variables are proxies for the correct-making features of the option, knowing which proxies drove the decision does not tell us whether the correct-making features of a decision are present. For Obermeyer's case, this means that we have to firstly work out whether medical need is, all things, considered the right criterion or if not, what the other correct-making

[8] Schmidt and colleagues argue that not knowing how the primary algorithm reached its decision can undermine knowledge by making the resultant belief unsafe even if justified and true (Baum et al., 2022; Schmidt et al., 2025).

factors are and how they relate to expected expenditure. If it turns out that we should distribute according to medical need, then it turns out that Obermeyer's algorithm makes incorrect decisions when it comes to black patients. This would be because even though the algorithm estimates expected expenditure accurately and unbiasedly, for a given level of expenditure, black patients have a much higher disease load. The algorithm thus underestimates medical need for black patients. If, on the other hand, we ought to distribute according to the expected benefit, the fact that black patients consume less healthcare, even when insured, when compared with white patients with the same disease load, is important. If this turned out to be true, then expected medical expenditure would be a reliable and unbiased proxy of expected benefit. Notice here that even here, expected expenditure even if a good proxy, is still a proxy for the ground reality we want to track, namely expected benefit. The basic point here is that, even if an algorithm reliably makes correct decisions, it is unlikely to base that decision on the correct-making features themselves rather than some proxy of those features. This matters because the only additional information that one gains from a transparent rather than opaque algorithm is about what the proxy variables are and how their values are calculated.

Defenders of algorithmic explainability and transparency might argue that knowing how the algorithm reached its decision can help understand whether the decision is correct (Baum et al., 2022; Schmidt et al., 2025). They argue that upon learning that the algorithm based its recommendation on the correct normative reasons, they can understand that the decision is correct. However, if the basis of the decision differs from the normative reason, the recommendation should be to reject the decision. However, information about how a decision was made, as we will shortly show, is not particularly useful to knowing whether a decision is correct and understanding why. Briefly, this is because it may not be immediately obvious what the correct-making features of a decision are and which option possesses a favourable balance of such features. That is to say, people are not perfect reasoners and cannot, without assistance, always tell when the normative reasons favour one option over another even if they possess all the relevant information. Furthermore, even very reliable algorithms rarely make their decisions on the basis of those correct-making features rather than some proxy for them.

To illustrate, consider the following case:

> Chemo: Suppose a patient, Dave, has been presented with a diagnosis of late-stage cancer, and further had only a few months to live. It turns out that while a course chemotherapy and radiotherapy could reduce the tumour size and potentially extend his life by a few years with some possibility of remission, he would likely suffer from various side effects including loss of appetite, nausea and various blisters and ulcers on the mouth and tongue. With these side effects, Dave will lose enjoyment of food. However, Dave comes from a food-centred cultural background and sees little point in living if he does not get to enjoy the foods that he likes and partakes with his community. As such palliative care is, all things considered, more in alignment with Dave's values and priorities than chemotherapy. However, Dave is not able to see this clearly. When discussing treatment options with his physician, Susan, they consult a decision support AI[9] which recommends palliative care. This contrasts with Susan's initial inclination, prior to consulting the AI, to recommend chemotherapy. It turns out that one of the factors that contributed to the AI's recommendation was the fact that Dave was a member of a particular ethnic group.

We can imagine that this would be because a number of the correct-making features in Dave's situation from his preferences to his susceptibility to side effects correlate with ethnicity. If the algorithm was justifiable, he would have learned the reasons why palliative care was best for him. If the algorithm was merely explainable, he would only have learned how the algorithm reached its particular recommendation without learning why the decision was best for him. Learning what an algorithm's proxy variables are and how they are calculated is often irrelevant to understanding whether and why a decision is correct because it is not always clear what the correct-making features of the decision are and it may be hard to notice if they are present.

To see why, consider the case where the AI is merely a black-box AI without a secondary AI. Susan and Dave know that it is designed to maximise Quality Adjusted Life Years (QALYs) and that it is highly reliable, but not how the algorithm calculates its results. Given that Susan disagrees with the algorithm's recommendation, how is she to respond to this disagreement? Given that Susan knows that the algorithm is more reliable than her, we might be inclined to say that she should defer to the algorithm. However, things are not so simple.

Suppose that Susan goes back over her reasoning for her decision and does not find anything that *she can identify* as wrong with her reasoning. Susan's reasoning is that while chemotherapy would likely result in blisters in Dave's mouth and loss of appetite, these are typically regarded as moderate and an acceptable price to pay for an increase in lifespan. Her reasoning is mistaken as she fails to account

[9] See IBM's Watson for Oncology for an example of such a decision-support AI (Jie et al., 2021).

for the fact that Dave cares about food. However, despite going over her reasoning again, she fails to identify that she missed out that Dave cares greatly about food for cultural reasons. If, contrary to the case, Dave did not care so much about food, then her reasoning would thereby be impeccable. After all, she would not have missed out on a crucial fact she had access to, namely that Dave comes from a cultural background that places great importance on food. If all other facts remained the same and the AI had given the same recommendation, the conditional likelihood that the AI's recommendation is correct given that she reasoned impeccably would be very low. This means that in the case where her reasoning is mistaken in some way but the mistake has not been detected by her, there is no reason that *she is aware of* to think that she, rather than the AI is mistaken. That is, she lacks guidance. Notably, here, mere epistemic justification is not enough. For instance, design publicity (Loi et al., 2021) provides epistemic justification by having a track-record of reliability and showing that the algorithm still operates in the same reliable way as it did in establishing the track-record. Even with this information, while she can appreciate that there is a high prior unconditional probability that the AI is *not* mistaken, she has less insight into the likelihood the AI is mistaken *given that she disagrees*. If she in fact reasoned impeccably, the conditional probability of the AI being mistaken (or at least unsupported by all the evidence) *given* that she disagrees is 1. If she did make a mistake, that probability is much lower. As such, she needs to be made aware that she overlooked the fact that Dave comes from a food-centred culture. Put more generally, she needs to be made aware of the incorrect-making features of her recommendation and, correspondingly, the correct-making features of the AI's recommendation.

However, algorithmic transparency provides the wrong kind of information to bridge this gap in guidance. Since the algorithm's variables tend to represent proxies of the correct-making features of the decisions rather than the correct-making features themselves, transparency will only reveal how the algorithm happened to calculate its answer. Unless, users like Susan know how the proxies relate to the correct-making features of the situation, she will, after learning how the algorithm arrived at the decision, not be any closer to understanding why the algorithm's decision is correct. In fact, Dave would make the *wrong* decision if he were to reject the algorithm's recommendation and go for chemotherapy just because the algorithm happened to base its recommendation partly on his ethnicity.

By contrast, if the algorithm was justifiable, Susan would be made aware of why palliative care is better for Dave. This would have made the flaws in her earlier reasoning apparent to her thus bridging the guidance gap. To be sure, the algorithm does not know idiosyncratic information about Dave that Susan lacks access to. After all, Susan does in fact have access to the fact that Dave comes from a particular background. It is she who typed it into the doctor's notes and made it available to the algorithm. Rather, Susan's mistake lies in, perhaps understandably, overlooking this fact. It is difficult to attend to all the relevant evidence at the same time (Booth and Peels 2010; Peels and Booth 2014; Podgorski 2016a, b). The algorithm merely draws Susan's attention to what she overlooked in her reasoning. Importantly, since algorithms only occasionally use variables which directly represent the ground truth, the justification will often not reflect how the algorithm actually arrived at the decision. Since the justification, by itself, is sufficient to provide guidance and further often does not reflect the inner-workings of the algorithm, transparency is neither necessary nor, in most cases, conducive to providing guidance for the user.

Justifiability, as we have seen, is valuable because it helps guide decision-making, especially in cases where the algorithm's recommendation differs from the user's own unaided judgment (Muralidharan et al., 2024). Improving someone's understanding of why a decision is correct can also increase the motivation to act accordingly (Colby, 2002; Deniz et al., 2021). Algorithmic transparency is distinct from justifiability because it is neither necessary, sufficient nor necessarily conducive towards helping the user understand why a given output of an algorithm is correct (if it is indeed correct). Nevertheless, it could still be the case that algorithmic transparency is required for reasons other than the guiding decision-making. Some cited reasons for algorithmic transparency are for procedural justice, in order to enable planning, in order to provide a basis for challenging AI decisions (Selbst & Barocas, 2018) and for trust. In this paper, we argue that either these considerations are misconceived or they support algorithmic justifiability instead of transparency.

Procedural justice

The concern for procedural justice is rooted in concerns about autonomy and respect for persons (Selbst & Barocas, 2018, pp. 1118–1119). The crucial worry is that decisions that affect us should not be arbitrary from our own point of view. There are two ways we might interpret this requirement. In the first sense, this requirement is straightforwardly a demand for justification. After all, when we are able to understand why a decision is correct, it cannot be arbitrary from our own point of view. Alternatively, the worry about arbitrariness might be cashed out in terms of not being based on arbitrary or irrelevant criteria. The worry is that because Justifiable AI provides a justification that does not match how it actually arrived at a decision, it might engage

in ethics washing. Consider the following case outside the medical context:

> Bank Loan: Sam, a black man, with a credit score of 680, applies for a loan at a bank where loan applications are processed by a justifiable but not transparent algorithm. His loan was rejected. Suppose that the primary AI was racist and tended to reject black applicants at higher rates than similarly situated white applicants. However, the secondary AI might give a reasonable justification for each particular decision. In Sam's case, the justification that was given was that his credit score was below 720, and his income was within a certain bracket. In fact, if this standard was applied uniformly, it would be reasonable. However, it is applied in a biased way. Similarly situated white applicants are typically rejected only if their credit score is 660. If this standard were applied uniformly, it would be reasonable too. Different thresholds merely reflect different reasonable attitudes towards financial risk. Any individual banker who sees the AI's decision may not necessarily recall the justification given for other similarly placed applicants. In each case, the banker concurs with the AI's decision because both standards are reasonable. The justifiable AI thus gives a seemingly reasonable justification for each decision but this ends up obscuring the underlying bias (Aivodji et al., 2019; Turpin et al., 2023). It would seem that we would want to know if the primary algorithm's output actually depended on illicit criteria.

However, while it is true that using the algorithm would be procedurally unjust, it does not follow that algorithmic transparency is the relevant remedy. Firstly, algorithmic transparency does not make it easier to detect procedural injustice. Secondly, it is neither necessary nor sufficient to detect procedural injustice.

With regards to the first point, defenders of transparency might think that algorithmic transparency is helpful in detecting procedural unfairness. To explain, in order to detect objectionable bias in the algorithm in Bank Loan, we need to conduct regular audits of the decisions to ensure that they are not made on mutually inconsistent criteria. By contrast, or so the argument goes, regular audits are not necessary with transparent algorithms because the existence of procedural justice or injustice can simply be read off the source code of the algorithm.

However, the claim that transparent algorithms do not require regular audits is implausible. Suppose that the Banker's AI is explainable and justifiable. Just as the fact that the primary algorithm relied on ethnicity in the Chemo case did not reveal whether the decision was justifiable to the patient, procedural unfairness cannot be simply read off from the explanation of any given case. One would, at the minimum, be required to look at multiple decisions over time and see whether there was a systematic pattern of requiring more stringent standards for black applicants than for white applicants.

The defender of transparency might reply that here the problem is that explainable models are insufficiently transparent. If we were to use white-box models instead, we would be able to survey the whole algorithm and see if it penalises black applicants simply for being black. However, this would be too quick. It assumes an overly simple model of what white-box algorithms are. A white-box model is not an explicitly programmed formula. Rather, white-box models still involve training the algorithm on a dataset. The difference with black-box models is that white-box models use simpler and more understandable statistical methods while black-box methods use more computationally sophisticated ones which resist analysis. It is the use of statistical methods (whether simple or otherwise) in the context of existing background injustice which results in algorithms which use race to determine the outcome. This means that given background racial injustice, it is almost inevitable that algorithms, even procedurally fair ones will use race as a factor in arriving at decisions.

To see how this could happen, consider a case where a banker uses an algorithm to estimate the likelihood of loan default. While this algorithm is not perfectly accurate, it is quite accurate and does not systematically overestimate or underestimate likelihood of loan default for members any given race (or sex). Intuitively, the process by which the banker makes the decision seems procedurally fair even if not entirely substantively fair[10]. For example, suppose that robust evidence suggests women are safer drivers than men (Aldred et al., 2021). It would be fair to provide lower insurance premiums to them. However, when there is background racial injustice, affected racial minorities will tend to have lower educational attainments, income, wealth, credit score, higher incarceration rates and health problems *as a result* of the injustice. All of these are indicators about a given applicant's likelihood of repaying the loan. This means that when calculating likelihood of repaying a given loan amount, there will be some association between the applicant's likelihood of repaying the loan and their

[10] Some might argue that since many of the factors affecting likelihood of default were worsened by historical and ongoing race and sex-based oppression, making loan decisions on the basis of likelihood of default exacerbates unjust inequalities and may be substantively unjust. Substantive fairness would, on this view, require some kind of affirmative action. See also work by Muralidharan, Savulescu and Schaefer (Muralidharan et al., 2025) mounting objections to calling such algorithms unfair.

race[11]. This will be the case even after accounting for all the *known* linkages between background injustice and ability repay loans. Moreover, not even all known linkages could be accounted for computationally. For instance, historical and structural injustices have resulted in different savings[12] and investment behaviour between black and white Americans, even after adjusting for income and education (Choudhury, 2001). Black Americans are more likely to have less diverse, lower yield portfolios (Choudhury, 2001). These different behaviours will in turn contribute to differences in wealth and their ability to repay a loan. While existing investment portfolios might accurately predict the investment behaviour of experienced investors, it does not do so for first-time investors. If there is no way of computationally isolating some of the known predictors of investment behaviour, then proxy variables like race and sex may be the only way to capture these contributions to the likelihood of repaying a loan. The unknown, or otherwise non-isolatable, connections will mean that functions that include the applicant's race (or sex) among other factors will still be more accurate than any function consisting only of the known factors that affect likelihood of repayment. This means that insofar as likelihood of repaying a loan is a procedurally fair criterion for deciding to issue a loan, even procedurally fair algorithms which issue loans on the basis of likelihood of repayment will base the decision to some degree on race. That is, procedurally fair algorithms, in calculating the likelihood of repayment, will use race as one of the variables. Even if the applicant's race is not explicitly encoded into the data, other proxies for race like the applicant's name, area code, etc. may be used.

At this point, one might think that if the algorithm chooses, at least partly, on the basis of race, it is not procedurally fair after all. One might think that procedural fairness requires using similar criteria for similar cases and in this case, the algorithm uses different criteria for applicants who differ only by their race. While procedural fairness does require treating similar cases similarly (Gillespie, 1975; Hart, 1958; Leventhal, 1980; Winston, 1974), it is not clear whether the cases are indeed relevantly similar here. Consider that the likelihood of repaying a loan is a relevant criterion for issuing one. After all, if banks indiscriminately lent money to those unlikely to pay back, they would eventually run out of money to lend. Since race, or some other proxy for race (like name or address[13]) would contain information important for predicting likelihood to repay, two otherwise similar people of different races are not similar in *all* relevant respects. This means that even if likelihood of repayment is unevenly distributed among racial groups, the centrality of repayment to the practice of lending money can justify seemingly different treatment. This might still strike us as unfair. People should not be penalised, when applying for loans, on the basis of their race. This is indeed true, but all this means is that issuing loans on the basis of likelihood of repayment is not *substantively* fair. Procedural fairness alone, especially given background injustice, does not suffice to establish substantively fair outcomes.

As such, the mere fact that an algorithm uses race as a variable does not mean it is procedurally unfair. If this is right, procedural fairness or unfairness cannot be read off an algorithm even if the algorithm were fully transparent. Therefore, procedural fairness can only be ensured by external monitoring measures. This would include regular audits to check whether the justifications for various decisions are consistent with each other, as well as enforcing transparency on the training data, validation protocols etc. None of this need involve finding out the internal workings of the algorithm.

This brings us to the second point as to whether algorithmic transparency is necessary and/or sufficient to detect procedural injustice. The previous discussion established that, strictly speaking, algorithmic transparency was not sufficient, in and of itself, to detect procedural injustice. That said, we might ask a slightly more nuanced question: could the regular monitoring of the internal workings of the algorithm together with the training data, validation protocols etc. be necessary or sufficient to detect procedural injustice? People might say that it could be at least necessary and often sufficient for the following reason: Procedural justice is a matter of basing one's treatment of relevantly similar cases on the same criteria. Algorithmic transparency, on this view, is necessary because we need to know the basis of the algorithm's decisions in order to detect whether it had used different criteria for different people who were relevantly similar to each other. Arguably it could also be sufficient in detecting procedural injustice if, when monitoring over time, we used it to detect whether it was basing its decisions on the same criteria for relevantly similar persons.

However, this argument for the necessity of algorithmic transparency for detecting procedural fairness presupposes a mistaken account of procedural justice. On this account,

[11] Neelakantan (2023), for instance, says that gap may be explained by income and net worth, but doesn't say whether all of the default gap is explained by these two variables.

[12] Notably, this does not refer to the savings rate. Differences in black and white savings rates among Americans are almost entirely accounted for by the interaction between one's income and one's neighbours' income (Charles et al., 2009).

[13] Importantly, even if laws like Title VII in America forbid explicitly encoding race into the dataset (Barocas & Selbst, 2016), a sufficiently rich dataset would allow algorithms to predict race (Gichoya et al., 2022) even if not explicitly encoded and thereby seemingly treat otherwise similar people of different races differently.

procedural fairness requires that the *algorithm base* its decisions on the same criteria for relevantly similar cases. However, there are, arguably, at least two objections to this view. Firstly, it is not the algorithm's decision-making process that matters but at most the (human) decision-maker's process instead. Moreover, the two can come apart. To see why, consider the following variant of the Bank Loan case.

> *Fair Bank Loan*: Shyam, an Indian man with a credit score of 680 applies for a bank loan and is rejected. The banker who made this decision did not use an algorithm and justifies this decision on the grounds that the credit score was less than 700. Looking at the recent records of the bank's decisions shows that the bank has consistently applied this same standard across similar cases.

Quite plausibly, we can regard the decisions made in Fair Bank Loan as procedurally fair. Suppose that we were to now modify Fair Bank Loan by adding a justifiable algorithm that presents the same consistent standards for all relevantly similar cases. That is, for everyone of similar income, education and level of assets, a credit score of less than 700 results in a rejection of the loan application. If the banker based her decisions on these standards we would still regard those decisions as procedurally just. This is because in both the fair cases, with or without the algorithm, the banker is making the same decisions for the same reasons. Suppose that we were to later find out that the algorithm used inconsistent criteria for its recommendations despite specifying a different consistent criteria as its justification. For instance, the actual threshold for Indian applicants was 710 instead of 700, It turns out that through a quirk of fate, none of the Indian applicants had a credit score between 700 and 710. It seems arbitrary to regard the decisions as being procedurally unfair. After all, as far as the decision-maker is concerned, the decisions made and the procedures are the same in all relevant aspects. The mere fact that in one case, her reasoning is prompted by an algorithm and, in another it is not, is not a relevant difference. The standards, the decisions and the reasoning employed are the same in both cases. The key point here is that even if the algorithm itself is unfair, the procedure by which the loan decisions were made is not, and it is the latter which matters, not the former. The allocation of loans in the real world can be justified.

One might worry that procedural fairness seems to, oddly, depend on the absence of cases of a particular kind. If an Indian applicant with a credit score of 705 was rejected by the algorithm, would the addition of this one case not render the whole process procedurally unfair? This seems objectionably odd. However, while this might cause the algorithm itself to be procedurally unfair, it does not mean that the process as a whole is itself procedurally unfair. After all, if a justifiable algorithm rejects an Indian applicant whose credit score is above 700, there are two possible kinds of justifications it might offer. Of the first kind, the justification it presents accurately represents the applicant's credit score. In that case the banker should be able to notice that the algorithm's decision is anomalous. In the second kind, the algorithm may "hallucinate" a false credit score. However, here, there would be a discrepancy between the stated score and the independently calculated score or other credit information provided. Even here, this allows the discrepancy to be detected. Either way, the banker could reject the algorithm's suggestion and issue the loan. As such, the loan issuance procedure would still be fair even if the algorithm which assists it is not as long as humans are able to justify the decision with reasons.

Secondly, even requiring only the human decision-maker to actually base their decision on the same criteria might be too strong a requirement. After all, we do not typically, when AI is not involved, try to mindread decision-makers (Zerilli et al., 2019). Instead, when we want to check whether decision-making was procedurally fair, we ask only whether the stated rationales for decisions were reasonable, applied consistently, and whether the decisions were consistent with those rationales. This is precisely what justifiable AI delivers. If we are to hold AI assisted decision-making to the same standards as we hold unassisted decision-making, it is algorithmic justifiability, not algorithmic transparency that is required.

If the above is right, then the internal workings of algorithms are the wrong facts to be tracking or auditing if we want to detect procedural (un)fairness. Hence, algorithmic transparency is neither necessary nor sufficient for detecting procedural (in)justice. By contrast, algorithmic justifiability may be at least as good, if not better, at detecting procedural injustice. Regular monitoring and auditing the justifications provided by a justifiable algorithm for consistency and soundness can help avoid ethics washing.

Planning to avoid bad outcomes

The thought behind the second rationale for transparency is more pragmatic: An algorithm may very well be unfair in various ways. Yet, knowing, in advance, how it is going to decide will help an agent know what to do to avoid bad outcomes (Selbst & Barocas, 2018, pp. 1120–1122; Wachter et al., 2018).

Our response is that where knowing how an algorithm makes decisions can provide guidance, algorithmic justifiability is even better. After all, any imperfect algorithm may sometimes decide against some agent on the basis of some

immutable characteristic. For instance, an algorithm which decides whether to issue a loan may recommend refusal because the prospective borrower, Sam, happened to be black. In this case, there is nothing that Sam could do to not be black. Even if race was only one consideration among others like income and currently possessed assets, there is still little that Sam could do. Sam could not, after all, in the immediate near term increase his income or purchase property in order to make up for the degree to which he is penalized on the basis of his race. After all, at this point in time, he is only taking a loan because he needs money! Hence, the thing that needs to be done to get around the algorithm can sometimes be impossible or infeasible.

However, a decision-maker who uses justifiable AI will reliably make decisions that everyone can see meet the appropriate normative standards. The banker, when using the justifiable AI may see, for instance, that given Sam's credit score, income and assets, there is no good justification for refusing him the loan. If this is right, it would be no harder to get around the algorithm by satisfying reasonable standards of behaviour instead of idiosyncratic ones drawn from the source-code of an imperfect algorithm.

One might worry that an algorithm which was merely justifiable would give the wrong advice about how to elicit a different response from the AI (Babic & Cohen, 2023). This is indeed true. However, eliciting different responses from the AI is valuable only if it is instrumental to eliciting different responses from the decision-maker. After all, it is ultimately the decisionmaker's choice which directly affects the well-being of the applicant or patient. With justifiable AI, the decision-maker does not merely defer to algorithm. She uses the provided justification to determine how to respond to the AI's output. In addition, as argued in the previous section, given justifiable AI, algorithmic transparency would not provide any further guidance to the decision-maker in deciding whether to accept or reject the AI's recommendation. As such, there is no point in knowing how to manipulate the AI's output if one already knows what to do in order to get a favourable decision from the decisionmaker.

In fact, since reasonable standards of behaviour are less likely to require the impossible[14], algorithmic justifiability is superior to transparency in terms of helping people navigate the system. To illustrate, all Sam needs to do in order to secure a loan from a banker using justifiable AI is keep a reasonably good credit score and work in, to the best of his ability, a sufficiently remunerative job. No matter how difficult this is for Sam, this is still significantly easier than doing the above *and* not being black.

[14] The point is not that the output of justifiable AI is always actionable. Rather, the point is that the justifiable AI's output is at least as actionable as transparent AI's. Sometimes this will be not actionable at all.

Challenging decisions

The third motivation for transparency is being able to challenge the decision. Users may challenge AI output if they knew how that output was reached. However, this is a mistake. Knowing how the output was reached has no bearing on whether the decision meets appropriate normative standards. This is because the question of whether a decision meets appropriate normative standards is fixed by the balance of correct and incorrect-making features of all the available options. These are fixed by the way the world (and the patient) actually is and the evidence available to decision-makers. The primary algorithm can at most pick out the option with the best balance of correct-making and incorrect-making features. It is only under unusual circumstances that an algorithm could change what makes a decision correct.

For instance, Dave, from the Chemo case knowing that the algorithm actually recommended palliative care in part because of his race has no basis for challenging the decisions because the decision, at the end of the day, is still the best one for him given his own values. Likewise, Sam from the Bank Loan case cannot challenge the algorithm's decision for being procedurally unfair *just because* the algorithm happened to base its recommendation partly on the applicant's race. What matters is whether there are consistent justifications given for the bank's decisions. Sam in the Bank Loan case can rightly challenge the decision only because the cited justifications for different decisions about similar cases cannot be made consistent with each other.

More generally, challengeability, rightfully conceived, is a virtue only insofar as decisions can be challenged for failing to meet appropriate normative standards. Decisions which demonstrably meet those standards should not be challenged. Therefore, since information about whether an algorithm's decision meets the relevant normative standards is provided by justifiable, but not transparent algorithms, challengeability requires algorithmic justifiability, not algorithmic transparency.

At this point, one might object that we have been too quick in dismissing the possibility that the internal workings of the algorithm could be relevant. Dave, upon learning that the algorithm's decision was partly based on race might change his priorities because he learned how the algorithm came to its decision. Upon changing his priorities, he might challenge the algorithm's decision on the grounds that it no longer fits his commitments and values. In that case, transparency would seemingly have helped Dave challenge the algorithm's decision by revealing how it came to the decision.

However, the role played by Dave coming to know how the algorithm made its decision in Dave's change of

priorities is merely incidental. Given that it is already clear how the decision fits Dave's priorities and values, the particular way in which the algorithm made its decision is not a good reason to change said priorities. Dave's changing his priorities in this way would be arbitrary, no different from changing his priorities because of a coin toss. The mere fact that Dave could possibly change his priorities on the basis of such arbitrary criterion does not mean that he ought to have access to this information. In fact, the more likely that people are to react arbitrarily and irrationally to some information, the more reason we have to *not* reveal such information. In general, there may be all sorts of reasons, good or bad, a person could have to change their priorities. The role of a decision-support AI is not to *change* what makes a decision correct; only to help identify the correct option and give reasons why. Hence, even if there is some information that could change the correct-making features of a given decision, there is no good reason to provide that information.

To sum up, defenders of transparency have a dilemma. Either knowing how the algorithm reached its decision does not change the patient's priorities or it does. If it does not, then any information thereby revealed (insofar as it is distinct from the justification) is irrelevant for challenging the decision. If the patient's priorities would be changed, the information is still not relevant because changing one's fundamental priorities in response to such irrelevant information is irrational and arbitrary. In either case, the algorithm being justifiable is sufficient.

Trust

> *They just don't trust what they can't explain* (Collins, 1999).

The link between explainability and trust seems almost platitudinous. This platitude has been leveraged to argue for algorithmic transparency. Interpretability and explainability have often been defended on the grounds that it is necessary for trustworthiness. The guiding intuition here is that you cannot trust an algorithm if you do not know how it works (Gilpin et al., 2018; Ribeiro et al., 2016; Rudin, 2019; Zerilli et al., 2022). However, we need to be careful about whether the kind of explainability or transparency required for trustworthiness is algorithmic transparency.

A further complication is that it is not altogether clear that AI, at the current state of technology, can be trustworthy. As of the moment this paper is being written, AI models are not moral agents in any meaningful sense[15]. However, on many accounts of trust (Asan et al., 2020; Baier, 1986; Darwall, 2017; Faulkner, 2017; Hinchman, 2017; Holton, 1994; Jones, 1996, 2017; McGeer & Petit, 2017; Muralidharan et al. forthcoming; R. M. Simpson, 2017; Simpson, 2012), moral agency is required for trustworthiness. On such views, to be trustworthy is to be more than merely reliable, it requires also having good will to the truster, or the ability or willingness to respond to the trust placed on them. Hence AIs cannot be trustworthy even if they are made transparent. Moreover, transparency does not increase reliability either[16]. On other accounts of trust (Ferrario et al., 2021; Kaplan et al., 2021), an AI might be considered trustworthy just in case they reliably perform the function they were designed for. Since transparency does not increase reliability, on these latter accounts, transparency does not increase trustworthiness.

Algorithmic justifiability, on the other hand, can improve the reliability of decision-making. If the primary algorithm makes a mistake, the secondary algorithm's attempt to justify that mistake would more likely involve some invalid reasoning or a citation of spurious evidence. This would be because any attempt to justify the unjustifiable must involve some mistake in fact or reasoning[17]. LLMs have also been found to hallucinate more when prompted to defend or account for falsehoods (Huang et al. 2023). If this is right, then, a well-trained justifiable AI would be less likely to hallucinate or generate *non-sequiturs* when the primary algorithm has not made a mistake than when it has. If the AI's user, qua subject-matter expert, detects the spurious evidence or invalid reasoning, that will improve the reliability of decision-making. Algorithmic justifiability thus improves the rate at which the user correctly detects that the primary AI's decision is correct.

This leads us to the question of whether algorithmic transparency is required for trusting people and systems which use AI. Regarding the claim that algorithmic transparency is necessary to trust persons, there are two things we might say.

Firstly, supposing that interpersonal trust requires transparency of some kind, the only kind of transparency that this could be is either transparency about decisional criteria and/or transparency about motivations. We might, for instance, be unwilling to put ourselves at the mercy of others if we are not certain about what criteria they use to make decisions or what considerations will actually motivate them to

[15] For a contrary view, see Railton (2022).

[16] According to some views, transparent models are less reliable than black-box models (Babic et al., 2021; Durán & Jongsma, 2021). However this is disputed by Rudin (2019). At the very least transparent models are not more reliable than Black-box models.

[17] Mistakes in reasoning can refer to both the making of an invalid inference and a failure to make valid inferences about the subject matter under consideration (Anantharaman, 2015).

act. Given the way in which algorithmic justifiability can guide decision-making, the criteria for the decision and the motivations for acting will usually be determined by the justification supplied by the justifiable algorithm. While the algorithm itself may have arrived at the decision on the basis of very different considerations, insofar as the person using the AI is not malicious, they will base their decision on the arguments presented by the secondary AI. Justifiable AI thus leaves us no less informed about the intentions and motivations of the AI user than explainable or interpretable AI.

Secondly, since algorithmic justifiability, not transparency is what enables physicians and patients to critically reflect on the output of the AI, only the former aids decision-makers in taking responsibility for the AI-assisted decision. Intuitively, it seems unreasonable to hold physicians responsible for the AI decision, if they had no way of knowing whether to accept or reject it (Kemper & Kolkman, 2019). Also plausibly, a person is trustworthy only to the extent they are willing and able to bear responsibility for their actions. Consider two equally reliable mechanics. One of them is willing to take responsibility for repairing your car and give you your money back if not repaired and the second is not. We judge the first to be more trustworthy than the second simply because she is willing to take responsibility while the latter is not. Likewise the system consisting of the Physician and Justifiable AI is more trustworthy than a system consisting of the Physician and interpretable or explainable AI because the physician in the first system is better able to take responsibility for the decision than the second. Hence, all things considered, if we care about trustworthiness, we should pursue algorithmic justifiability instead of algorithmic transparency.

Conclusion

Summing, up, algorithmic justifiability, which is distinct from algorithmic transparency, is valuable because it can help guide decision-making and also helps motivate acting according to the decision that was made. We have also seen that algorithmic justifiability, if present, can be expected to do at least as well as transparency in achieving procedural justice in decision-making and planning for bad decisions. In fact, algorithmic justifiability, where present, makes algorithmic transparency unnecessary. Moreover, we have also seen that it is a mistake to suppose that algorithmic transparency is either necessary or even conducive towards helping affected parties challenge AI-aided decisions or trust AI users. By contrast, algorithmic justifiability is what we should pursue if we wanted AI-aided decisions to be challengeable or if we wanted AI-aided decision-makers to be trustworthy. If this is right, then decision-making AI should be designed with an eye towards justifiability instead of transparency. Likewise AI guidelines like the EU AI Act (European Parliament 2024) should require justifiability instead of transparency. It is AI justifiability, not transparency, that is ethically required.

Acknowledgements We'd like to thank Owen Schaefer for comments on an earlier version of this paper. We would also like to thank the reviewers for their many useful comments, the incorporation of which improved the paper.

Author contributions All authors contributed to the drafting of the manuscript. The first draft of the manuscript was written by AM. All authors commented on and edited previous versions of the manuscript. All authors (AM, and JS) read and approved the final version of the manuscript.

Funding statement This research was funded by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-GV-2023-012), the Singapore Ministry of Health's National Medical Research Council under its Science Health, and Policy Relevant Ethics, Singapore (SHAPES) Programme (MOH-000951) and NUHS Internal Grant Funding [Grant/Award Number: NUS Start-up Grant/NUHSRO/2022/078/Startup/13]. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests Julian Savulescu is a Bioethics Committee consultant for Bayer.

References

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: the risk of rationalization. *arXiv*. https://doi.org/10.48550/arXiv.1901.09749

Aldred, R., Johnson, R., Jackson, C., & Woodcock, J. (2021). How does mode of travel affect risks posed to other road users? An analysis of English road fatality data, incorporating gender and

road type. *Injury Prevention*, *27*(1), 71–76. https://doi.org/10.1136/injuryprev-2019-043534
Anantharaman, M. (2015). Defending the Uniqueness Thesis: A Reply to Luis Rosa. *Logos and Episteme*, *6*(1), 129–139.
Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *Journal of Medical Internet Research*, *22*(6), 15154. https://doi.org/10.2196/15154
Babic, B., & Cohen, I. G. (2023). The Algorithmic Explainability Bait and Switch. *Minnesota Law Review*, *108*, 857–909. https://ssrn.com/abstract=4541487
Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. *Science*, *373*(6552), 284–286. https://doi.org/10.1126/science.abg1834
Baier, A. (1986). Trust and Anti-trust. *Ethics*, *96*(2), 231–260. https://www.jstor.org/stable/2381376
Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732. https://doi.org/10.15779/Z38BG31
Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, *35*(1), 12. https://doi.org/10.1007/s13347-022-00510-w
Beijing Academy of Artificial Intelligence (2019). Beijing AI Principles. *Datenschutz und Datensicherheit*, *10*, 656. https://link.springer.com/content/pdf/10.1007/s11623-019-1183-6.pdf
Beisbart, C. (2021). Opacity thought through: on the intransparency of computer simulations. *Synthese*, *199*, 116443–111666. https://doi.org/10.1007/s11229-021-03305-2
Booth, A., & Peels, R. (2010). Why Responsible Belief is Blameless Belief. *The Journal of Philosophy*, *107*(5), 257–265. https://www.jstor.org/stable/25764445
Charles, K. K., Hurst, E., & Roussanov, N. (2009). Conspicuous Consumption and Race. *The Quarterly Journal of Economics*, *124*(2), 425–467.
Choudhury, S. (2001). Racial and Ethnic Differences in Wealth and Asset Choices. *Social Security Bulletin*, *64*(4).
Colby, A. (2002). Moral Understanding, Motivation and Identity. *Human Development*, *45*(2), 130–135. https://www.jstor.org/stable/10.2307/26763667
Collins, P. (1999). *You'll be in My Heart*.
Darwall, S. (2017). Trust as a Second-Personal Attitude (Of the Heart). In P. Faulkner, & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 35–50). Oxford University Press.
Deniz, S., Akbolat, M., Cimen, M., & Unal, O. (2021). The Mediating Role of Shared Decision-Making in the Effect of the Patient–Physician Relationship on Compliance With Treatment. *Journal of Patient Experience*, *8*, 1–5. https://doi.org/10.1177/23743735211018066
Durán, J. M. (2021). Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence*, *297*, 103498. https://doi.org/10.1016/j.artint.2021.103498
Durán, J. M. (forthcoming). Beyond Transparency: computational reliabilism as an externalist epistemology of algorithms. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*.
Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*, 329–335. https://doi.org/10.1136/medethics-2020-106820
Elgin, C. Z. (2017). *True Enough*. MIT Press. https://doi.org/10.7551/mitpress/11118.001.0001
European Parliament. The Artificial Intelligence Act (2024).
Faulkner, P. (2017). The Problem of Trust. In P. Faulkner, & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 109–128). Oxford University Press.
Ferrario, A., Loi, M., & Vigano, E. (2021). Trust does not need to be human: It is possible to trust medical AI. *Journal of Medical Ethics*, *47*(6), 437–438. https://doi.org/10.1136/medethics-2020-106922
Gaus, G. (2011). *The Order of Public Reason*. Cambridge University Press.
Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L. C., et al. (2022). AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, *4*(6), e406–e414. https://doi.org/10.1016/S2589-7500(22)00063-2
Gillespie, N. C. (1975). On Treating Like Cases Differently. *The Philosophical Quarterly*, *25*(99), 151. https://doi.org/10.2307/2217630
Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Lalana, K. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics* (pp. 80–89). Presented at the DSAA.
Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8
Hart, H. L. A. (1958). Positivism and the Separation of Law and Morals. *Harvard Law Review*, *71*(4), 593. https://doi.org/10.2307/1338225
Hills, A. (2009). Moral Testimony and Moral Epistemology. *Ethics*, *120*(1), 94–127. https://www.jstor.org/stable/10.1086/648610
Hinchman, E. S. (2017). On the Risks of Resting Assured: An Assurance Theory of Trust. In P. Faulkner, & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 51–69). Oxford University Press.
Holdren, J. P., Smith, M., Bruce, A., Felten, E., Lyons, T., & Garris, M. (2016). *Preparing for the Future of Artificial Intelligence*. Executive Office of the President, National Science and Tehnology Council, Committee on Technology.
Holton, R. (1994). Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy*, *72*(1), 63–76. https://doi.org/10.1080/00048409412345881
Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H. (2023, November 9). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv. http://arxiv.org/abs/2311.05232. Accessed 2 October 2024.
Jie, Z., Zhiying, Z., & Li, L. (2021). A meta–analysis of Watson for Oncology in clinical application. *Nature Scientific Reports*, *11*. https://doi.org/10.1038/s41598-021-84973-5
Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, *107*(1), 4–25. https://www.jstor.org/stable/2382241
Jones, K. (2017). But I Was Counting on You! In P. Faulkner, & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 90–108). Oxford University Press.
Kaplan, A. D., Kessler, T. T., Christopher, B. J., & Hancock, P. A. (2021). Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 1–25. https://doi.org/10.1177/00187208211013988
Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information Communication and Society*, *22*(14), 2081–2096. https://doi.org/10.1080/1369118X.2018.1477967
Leventhal, G. S. (1980). What Should Be Done with Equity Theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social Exchange* (pp. 27–55). Springer US. https://doi.org/10.1007/978-1-4613-3087-5_2
Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics and Information Technology*, *23*(3), 253–263. https://doi.org/10.1007/s10676-020-09564-w

McGeer, V., & Petit, P. (2017). The Empowering Theory of Trust. In P. Faulkner, & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 14–34). Oxford University Press.

Muralidharan, A. (2023). Political Liberalism and Reasonable Disagreement. *Social Theory and Practice*, *49*(1).

Muralidharan, A., Savulescu, J., & Schaefer, G. O. (forthcoming) (Eds.). Trust as a Moral Power. *Kennedy Institute of Ethics Journal*. https://philpapers.org/rec/MURTAA-12

Muralidharan, A., Savulescu, J., & Schaefer, G. O. (2024). AI and the need for Justification to the Patient. *Ethics and Information technology*, *26*. https://doi.org/10.1007/s10676-024-09754-w

Muralidharan, A., Savulescu, J., & Schaefer, G. O. (2025). Public Justification and Normatively Meaningful Bias: Against Imposing Egalitarian Accounts of Algorithmic Bias. *Bioethics*, 1–12. https://doi.org/10.1111/bioe.70047

Neelakantan, U. (2023). *Black-White Differences in Student Loan Default Rates Among College Graduates* (No. 23–12). Baltimore, Charlotte: Federal Reserve Bank of Richmond. https://www.richmondfed.org/publications/research/economic_brief/2023/eb_23-12

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.

Peels, R., & Booth, A. (2014). Why Responsible Belief is Permissible Belief. *Analytic Philosophy*, *55*(1), 75–88. https://doi.org/10.1111/phib.12036

Podgorski, A. (2016a). Dynamic Permissivism. *Philosophical Studies*. https://doi.org/10.1007/s11098-015-0585-z

Podgorski, A. (2016b). Dynamic Conservatism. *Ergo*, *3*(13), 349–376. https://doi.org/10.3998/ergo.12405314.0003.013

Railton, P. (2022, May 23). *Lecture 3: 2022 Annual Lectures in Practical Ethics*. Oxford, United Kingdom. https://www.practicalethics.ox.ac.uk/uehiro-lectures-2022#tab-3010806

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (pp. 1135–1144). Presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schmidt, E., Putora, P. M., & Fijten, R. (2025). The Epistemic Cost of Opacity: How the Use of Artificial Intelligence Undermines the Knowledge of Medical Doctors in High-Stakes Contexts. *Philosophy & Technology*, *38*(1), 5. https://doi.org/10.1007/s13347-024-00834-9

Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review*, *87*(3), 1085–1138. https://ir.lawnet.fordham.edu/flr/vol87/iss3/11

Simpson, T. (2012). What is Trust? *Pacific Philosophical Quarterly*, *93*, 550–569. https://doi.org/10.1111/j.1468-0114.2012.01438.x

Simpson, R. M. (2017). Permissivism and the Arbitrariness Objection. *Episteme*, *14*(4), 519–538. https://doi.org/10.1017/epi.2016.35

Smith, M. (1994). *The Moral Problem*. Blackwell Publishing.

Tan, B. O., Patrick, Ngiam, K. Y., Chin, J. J., Dunn, M., Kon, O. L., Krishnaswamy, P., et al. (2023). *Ethical, Legal and Social Issues Arising From Big Data and Artificial Intelligence Use in Human Biomedical Research*. Bioethics Advisory Committee.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. Presented at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans. https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf

Vallier, K. (2011a). Against Public Reason Liberalism's Accessibility Requirement. *Journal of Moral Philosophy*, *8*(3), 366–389. https://doi.org/10.1163/174552411X588991

Vallier, K. (2011b). Convergence and Consensus in Public Reason. *Public Affairs Quarterly*, *25*(4), 261–279.

Wachter, S., Mittelstadt, B. D., & Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Watson, D. S., & Floridi, L. (2021). The explanation game: a formal framework for interpretable machine learning. *Synthese*, *198*(10), 9211–9242. https://doi.org/10.1007/s11229-020-02629-9

Winston, K. I. (1974). On Treating Like Cases Alike. *California Law Review*, *62*(1), 1. https://doi.org/10.2307/3479821

Zednik, C., & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines*, *32*(1), 219–239. https://doi.org/10.1007/s11023-021-09583-6

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. https://doi.org/10.1007/s13347-018-0330-6

Zerilli, J., Bhatt, U., & Adrian, W. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*. https://doi.org/10.1016/j.patter.2022.100455