Minds and Machines (2026) 36:14
https://doi.org/10.1007/s11023-026-09768-x

Accuracy is not all you need! The Reasons to Require AI Explainability

Stefan Buijsman[1]

Received: 28 March 2025 / Accepted: 28 January 2026

Abstract
Do we need explanations of AI outputs in order to use AI systems (in high-risk settings)? This question has been actively debated recently, with one group denying that explanations are needed as long as the AI system is sufficiently accurate. What matters, according to them, is that outcomes improve. The other group argues that we have procedural reasons, centered around autonomy and self-advocacy, which trump outcome-based arguments to the contrary. I here present a set of arguments to show that outcome-based arguments should in fact also favor explainability for many of the current systems, as challenges with human oversight and accountability often lead to worse overall outcomes even if a more accurate AI system is integrated. Critics of explainability overlooked the fact that AI operates within a broader socio-technical system, and its accuracy alone tells us little of the final outcomes. In addition, I consolidate the procedural arguments and present a view of the upshot of these arguments. On this, we should avoid applications of AI that largely replace decision-making (relegating humans to the position of checking outputs). We can, however, use AI in other roles even for high-risk decision making while conforming to all of the requirements set by both outcome-based and procedural arguments. What matters, in the end, is the ability to explain decisions, and with the right role for AI that is possible even when supported by opaque systems.

Keywords Explanation · AI Ethics · Explainable AI · Right to explanation · Procedural justice

✉ Stefan Buijsman
s.n.r.buijsman@tudelft.nl

[1] TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

Published online: 27 February 2026

1 Introduction

To what extent do AI systems have to be explainable in order to use them? With the widespread introduction of Artificial Intelligence, including in high-risk settings such as healthcare (Shaheen, 2021), the military (Rashid et al., 2023) and the public sector (Wirtz et al., 2019), this is becoming a pressing issue. Many of the most advanced AI systems, with deep neural networks as the paradigmatic example, are (epistemically) opaque (Beisbart, 2021; Humphreys, 2009). That is, we lack explanations for why they produce the outputs that they do. While we do have access to the operations which these systems perform on the input data to reach the output, there are too many of them (making the system unsurveyable) and they are sub-symbolic (making it hard to interpret any individual calculation). As a result, we do know in general terms what the system is doing, but we lack explanations for why it produces one particular output rather than another, both at the individual input level and at the level of general behavior of the system. As a result of the latter, mistakes made by AI systems are hard to anticipate, as e.g. Tesla's self-driving car software incorrectly classified the full moon as an orange traffic light (Levin, 2021), making the cars slow down in the middle of the highway. Such mistakes are generally only spotted through extensive testing, as the systems are not explainable enough to understand their current shortcoming before observing them.

These challenges in explainability understandably raise the question: are we normatively justified to use these systems in high-risk settings where mistakes are costly despite their opacity? And if not, what kind of explanations should we require from these systems before they are admissible? These issues have been hotly debated in the literature, with roughly two types of answers emerging. On the one hand there are positive answers that yes, we can (and perhaps even should (Spencer et al., 2024)) use opaque AI systems even in high-risk settings (Durán & Jongsma, 2021; Krishnan, 2020; London, 2019). The details of these accounts are discussed in Sect. 2, but the shared sentiment here is that as long as the reliability of the systems is sufficient then we don't need explanations. While none of the authors frame it explicitly in this way, I argue that these arguments as essentially appealing to an outcome-based view of procedural justice: as long as a (decision-making) procedure leads to the right outcomes often enough, it is justified to use that procedure. Whether it is explainable or not is of secondary concern at best. While this is an intuitive way to think of our use of AI systems—why object to using an unexplainable system when it leads to better results?—I argue that this view misses two important aspects of our interaction with AI. These stem from the fact that AI systems rarely lead to decisions in isolation, but are instead part of a larger socio-technical system. In Sect. . 3.1 I argue that the interaction between an AI system and a human decision-maker is often difficult in the absence of explanations, and accurate but opaque systems can in fact harm the quality of the final outcomes. In section 3.2 I consider the use of accountability and contestability to improve a system over time, where likewise explanations are typically used (outside of the AI context) to ensure that the right changes are made to a decision-making process. In effect this gives a (qualified) argument in favor of explainability on the terms of those denying such a requirement.

To ensure we have the full set of arguments in favor of explainability I then discuss the rest of the literature in Sect. 4. This is needed to determine whether my proposal on how to move forward formulated in Sect. 5 satisfies all of the reasons for explainability. The other positive claims for a need of explainability have, namely, focused on the recipients of (high-stakes) decisions and the decisions themselves (Robbins, 2019), in contrast to my additional argument which focuses on the needs of decision-makers. Authors in this camp argue that we require explanations to ensure proper self-advocacy (Vredenburgh, 2022), legitimacy (Lazar, 2024), autonomy (Vaassen, 2022) and shared decision making (Chan, 2023). The challenge for these arguments, as well as my additional outcome-based argument, is whether these truly lead to a requirement for explainability. The explanations of AI systems that we can currently offer typically fall short of the stated goals (including mine for appropriate reliance and system improvement), so it seems that if we truly have a strong requirement that systems are explainable then we simply cannot currently use AI systems in high-risk settings. Instead, workarounds in terms of simpler types of transparency measures have been proposed for the goals of legitimacy and self-advocacy (Buijsman, 2024; Loi et al., 2021), and auditing might help to improve the system.

I therefore propose in Sect. 5 that a lack of explainability may not give an immediate reason to reject the use of an AI system. It does, however, give us a strong reason to look for alternative setups that allow us to meet more of the procedural requirements of decision-making as well as better socio-technical systems. What matters, on the requirement as presented here, is whether we can explain the final decision itself. Through careful design of human-AI interaction we can create socio-technical systems where humans can explain the decision without reference to the opaque AI, while still being supported by opaque AI in their decision-making. I argue that this satisfies the critics of opaque AI, while the AI system itself remains unexplainable. The explainability requirement argued for here thus does not rule out the use of opaque AI, although it does limit the uses to which we can put such systems.

2 Outcome-Based Critiques of Explainability

The arguments against a requirement that AI systems are explainable vary in their details, but the underlying idea tends to be the same: as long as an AI system is sufficiently accurate (e.g. as accurate as a human decision-maker in that context), then it can be used even when it is unexplainable. What matters is that the outcomes are no worse, and ideally better, thanks to the inclusion of the AI system in the process. Explainability can be a nice bonus, but reliability is what matters. This is especially explicit in the account of computational reliabilism (CR) presented by Durán & Jongsma (2021, p.332): "CR states that researchers are justified in believing the results of AI systems because there is a reliable process (ie, the algorithm) that yields, most of the time, trustworthy results." Medical practitioners would for example be "justified in trusting that a given dose for chemotherapy is right because dose-AI is a reliable medical AI system." Durán & Jongsma (2021, p.332) It is important to note here that computational reliabilism uses a fairly substantial definition of reliability: "A reliable algorithm is one that has been specified, coded, used, and maintained

utilizing reliability indicators. These reliability indicators stem from formal methods, algorithmic metrics, expert competencies, cultures of research, and other scientific endeavors." (Duran, 2024, p.1) This includes the technical accuracy of the system, but also incorporates the idea that the system outputs should be reliable indicators for the decision being made. For example, a system that outputs a criminality score based on images of faces may perform well on a technical accuracy test, but since expert knowledge suggests that facial structure is not a reliable indicator for criminality such an AI system would still fail to be reliable on CR. As long as this requirement of reliability is met, however, the system is deemed to be trustworthy and acceptable for use in high-risk settings. In their words: "CR becomes the solution to epistemic opacity in the most unusual way: it makes no attempts to solve it." (p.332 Durán and Jongsma (2021))

While (Durán & Jongsma, 2021) are unusual in the strong reading of reliability, the idea that reliability is what really matters is a common one. (London, 2019) reaches the same conclusion in the medical setting, that accuracy of the AI system is the only relevant consideration for their use. In this he voices the oft-heard argument that people are not perfect at explaining their decisions—and specifically that medical practice has often lacked a theoretical understanding of diseases and treatments that explained their efficacy—so why should we hold AI systems to a higher standard? His answer is that we shouldn't. Instead, "[r]ecommendations to prioritize explainability or interpretability over predictive and diagnostic accuracy are unwarranted in domains where our knowledge of underlying causal systems is lacking. Such recommendations can result in harms to patients whose diseases go undiagnosed or who are exposed to unnecessary additional testing." (London, 2019, p.20) The reasoning here is clear: less accurate systems lead to worse outcomes for patients. Since we have a duty to prevent harms to patients, we should therefore prioritize accuracy over explainability to ensure optimal outcomes.

Such an argument fits very well with outcome-based approaches to procedural justice. The best known of these is Rawls' notion of imperfect procedural justice (applicable here, given that AI systems are rarely perfectly accurate), where a procedure is just if it leads to the right outcome often enough (Rawls, 2001). Individual decisions are justified if they are the result of a just procedure that has been consistently applied. The typical illustration of this account is that of the judiciary: a trial procedure is just if it leads to the conviction of the guilty and the release of the innocent often enough. Practices such as allowing the accused access to an attorney, the impartiality of judges, the presentation of evidence, all help to achieve better outcomes of the procedure. Moreover, individual court decisions, even if they are not always correct, are justified because of their consistency with the just trial procedure, and often higher courts exist to test exactly that consistency. It is important to note here that the justification at issue is a normative justification. (Computational) Reliabilism can also be read as discussing epistemic justification, in which case the real-world outcomes are irrelevant to the epistemic question of justification. The focus in this paper is on normative justification (and so I make no claims that epistemic justification would require explainability): can we use opaque AI systems in high-risk settings? Epistemic justification may be a pre-requisite for doing so, but normative justification is certainly not the same. Translating this to the AI setting, accuracy is

the logical metric to focus on in order to ensure optimal outcomes and thus normative justification. If the system is more accurate, then (in line also with London's reasoning above) fewer mistakes will be made and thus outcomes should improve.

This may not cover all of the goals we have for AI systems, and in particular the goal of preventing discrimination. If ensuring the right outcomes is interpreted as ensuring just outcomes, then this would entail that outcomes need to be both accurate and fair. We may therefore require more than just technical accuracy. That consideration has not been overlooked, and Krishnan (2020) addresses it by considering three different reasons for requiring explainability. First, she discusses the justification problem: AI systems need to be explainable in order to provide justification to the decision-makers. Here the response is, similar to that of Durán and Jongsma (2021), to point to a reliabilist epistemology on which we acquire justification as long as the system/belief-forming procedure is sufficiently reliable. Second, she discusses the non-discrimination problem: explanations may be needed to spot cases of discrimination and prevent them. Krishnan's answer here is that fair machine learning techniques are now readily available (Carey & Wu, 2023; Caton & Haas, 2020) and do not require the model to be explainable. Access to outcome distributions is sufficient to identify biases, and there is also a range of techniques to mitigate and correct for biases that do not require AI systems to be explainable. We can therefore get to just outcomes without explainability.

The conclusion from this line of reasoning is that we should often opt for AI systems that are not explainable, especially in high-risk settings where we care about using just procedures. If it is correct that there is a trade-off between accuracy and explainability on which more accurate (complex) systems are less explainable, and more explainable (simpler to grasp) systems are less accurate (Holzinger et al., 2019) then the idea that accuracy is what matters will lead to the use of opaque systems. How strict of a trade-off this is is unclear, as empirical results fail to find clear inverse relations between (metrics of) explainability and accuracy (Bell et al., 2022; Dziugaite et al., 2020; Nesvijevskaia et al., 2021; Wanner et al., 2021). However, it does encourage the use of complex systems that perform as accurately as possible. Yet, as I argue in the next two sections, this is in fact no guarantee to optimal outcomes. What this focus on the accuracy of the AI system misses is that these systems rarely operate independently. Instead, they are part of a broader socio-technical system, and human decision-makers are in between the AI system and the final outcomes—especially in high-stakes settings. That limits the degree to which the accuracy of the system informs us of the proportion of just outcomes.

3 The Limits to Reliability

3.1 Appropriate Reliance

High-risk AI systems are unlikely to be implemented in a way that their output has immediate effects on people. Human oversight is a common requirement (e.g. in the EU AI Act; Enqvist (2023)) and is most visibly implemented by mandating that AI outputs are merely there for decision support, and that humans make the actual (final)

decisions. In such settings the accuracy of the AI system does not automatically translate into the correctness of the final decisions. That would only be the case if people appropriately rely on these systems, ideally even improving on their performance by correcting the system when it is wrong (but following it whenever it is right).

In practice appropriate reliance has proven to be very difficult. People under- or over-rely on AI systems, meaning that final outcomes can be no better or even worse than when humans make decisions on their own, even when the AI system is more accurate than they are. For example, Schemmer et al. (2023) conducted a study where participants had to decide whether a shown hotel review was deceptive or not. Initially they made a decision themselves, then were shown an AI output at 86% accuracy, to then be allowed to reconsider. Performance was incentivized by paying people additional compensation per correct answer, but to little effect. The performance in the control group was 55.5% without AI assistance, but only 53.9% with AI assistance. The simple reason was that in cases of disagreement, participants opted to stick to their initial answer 72% of the time. Although the AI system was a lot more accurate, that made no difference to the final outcomes because people didn't follow the AI outputs in their decision-making. Clearly, AI accuracy cannot be all that matters even when we accept the outcome-based reasoning of the critics of an explainability requirement.

How the accuracy of the AI system is related to the decision of human decision-makers remains a complex topic, with more questions than answers. Both the stated accuracy of the system (the percentage presented to decision-makers beforehand) and the observed accuracy of the system (actual performance over time) matter. Yin et al. (2019) found that in the simulated task format typical to these studies (where laymen complete a set number of tasks over a crowd working platform) the performance of the model certainly has an effect on reliance, but proper calibration is difficult. When participants were told that the model is 95% accurate, they agreed with the system 81% of the time, whereas on a stated accuracy of 60% this agreement was still 75%. Unfortunately the overall performance of the decision-makers is not reported, but the authors do give more details on how this reliance on the system evolves under different conditions. When getting feedback halfway, there is a clear difference depending on what the observed accuracy of the AI system is (how often it is reported to have gotten the right answer v.s. how often the participant made the right decision). If this observed accuracy is 80%, agreement with the AI system settled at 80–85%, whereas an observed accuracy of 55% led to around 75% agreement and an accuracy of 100% led to 85–90% agreement with the system (depending on the stated accuracy beforehand; the higher the stated accuracy, the higher the agreement with the system). People therefore do start to rely on the system more as it is experienced as being reliable and rely on it less if it is perceived as inaccurate. Yet, He et al. (2023) found no statistical difference between participant accuracy depending on different levels of AI accuracy (and under-performance compared to the AI system in all cases), which has been linked to participant over-estimation of their own performance (He et al., 2023).

On the other hand, people can also easily over-rely on AI systems. Klingbeil et al. (2024) used a game setup to test people's reliance on AI advice in (risky) financial settings and found that the "mere knowledge of advice being generated by an AI causes people to overrely on it, that is, to follow AI advice even when it contradicts

available contextual information as well as their own assessment. Frequently, this overreliance leads not only to inefficient outcomes for the advisee, but also to undesired effects regarding third parties." (Klingbeil et al. (2024), p.1)

The finding that people can start to overrely on AI advice, no longer effectively paying attention, can also be seen in more realistic settings. Dell'acqua (2022) ran a study with 181 professional recruiters to evaluate 44 job candidates, while being aided by an AI system. Against what one might expect, recruiters who were helped by a highly accurate system (85% accuracy) did worse overall than those helped by a less accurate AI system (75% accuracy). Dell'Acqua notes that this decrease in performance with the more accurate system is accompanied by a decrease in effort on the part of the recruiter. They spend less time and make fewer clicks when aided by a better AI system. This, in turn, was mediated by experience as "[r]egardless of whether experience comes from a background in HR or from a familiarity with AI technologies; greater experience leads to better performance when subjects are assigned to "Bad AI" [75% accuracy], and to worse performance when subjects are assigned to "Good AI" [85%] but especially "Perfect prediction" [100%]. Experienced recruiters were, in general, more likely to think independently and not to exclusively follow the AI's advice."(Dell'acqua (2022), p.28) In yet another setting, where 758 BCG consultants solved creative product innovation and business problem solving tasks. On product innovation either alone or aided by GPT-4, a somewhat similar dynamic regarding human effort was observed (Candelon et al., 2023). On product innovation, where Generative AI performs very well, overall individual performance of the AI-aided consultants improved by 40% compared to that of the control group (more ideas were generated) though at the expense of collective diversity (which was 41% lower). When it comes to business problem solving, however, where Generative AI misses contextual information, the answers provided by the AI-aided consultants were 23% worse than those of the unaided consultants. Here, too, a reduction of effort on the part of the consultant is likely to blame, leading to harmful overreliance as quality decreased even further for the consultants given a brief training on how to best use GPT-4 for these tasks (they showed a 29% decrease versus a 16% decrease for consultants using AI without training). Most likely, these consultants copied the AI output more readily, with less independent thinking, leading to worse performance as GPT-4 is not yet accurate enough on business problem tasks.

In short, for various reasons a more accurate AI system may not lead to improved outcomes. Both under- and over-reliance by human decision-makers can mean that outcomes do not improve or even deteriorate. In addition, especially in social and evaluative contexts, accuracy itself can be difficult to pin down (Robbins, 2025) in large part because the AI outputs often act as interventions in the socio-technical system (Liu et al., 2025). Since it is the final outcomes that matter, purely looking at the technical accuracy (or the reliability, as suggested by computational reliabilism) of an AI system is insufficient. Technical accuracy may not translate into the correctness of the final decisions. And even when people do take over the AI outputs directly, the accuracy/correctness measure may be ill-defined, in particular when AI outputs lead to interventions in complex social systems.

So do explanations make a difference here? On the face of it, the empirical literature on explainable AI suggests it doesn't: when participants are shown feature

importance scores along with AI outputs, or counterfactuals, their reliance on the system and overall performance shows little to no improvement (Hase & Bansal, 2020; Nagendran et al., 2023; Van Der Waa et al., 2021; Wang & Yin, 2021). At the same time, these studies also find that people perform no better on the common metrics for how well they understand the AI system, such as whether they are able to predict the output of the system for new inputs Chromik et al. (2021); Van Der Waa et al. (2021). Since most philosophical accounts of explanation consider that grasping an explanation allows someone to better predict a phenomenon, the more plausible reading of these studies is that explainable AI methods currently fail to reduce the opacity of these systems, and that therefore these studies tell us little about the effect of understanding on overall performance. It does, however, relate to the last argument consider by Krishnan (2020, Sect. 3.3). She discusses the 'reconciliation problem' of how to integrate human judgements and AI outputs, and suggests that interpretability is not needed as long as we know which features were considered by the AI system. That, at least, seems an unlikely solution given that the feature importance methods tested in these empirical studies fail to improve the situation.

So does the challenge of appropriate reliance present us with an argument for explainability? Despite limited evidence, there are good reasons to think that explanations will help. In a few studies where researchers used explanations tailor-made to the task in a way that made it easier to check the AI system, overreliance was reduced and overall performance improved (de Jong et al., 2025; Vasconcelos et al., 2023). Additionally, Spatola (2024) investigated the trade-off between efficiency and explainability when relying on AI systems. Participants in this study were asked to solve different reasoning puzzles, assisted by a chatbot that either gave only the answer or gave the answer as well as explanations. Those who only got the right answer relied significantly more on the chatbot than the participants who saw explanations. That became problematic as soon as the chatbot performance dropped: reliance on the AI remained the same despite the fact that the AI outputs were more often incorrect. The more participants relied on the system initially (when it was accurate), the worse their overall performance was due to the reduction of the system's accuracy over time. As a result, "while efficiency-focused AI solutions enhance immediate performance, they risk over-assimilation and reduced vigilance, leading to significant performance drops when AI accuracy falters." (Spatola (2024), p.1) Explanations seem to reduce that risk.

Furthermore, we use explanations in many of our high-stakes decision-making in order to allow others to evaluate that decision. By highlighting the reasons for the decision, it is easier to see whether it was made for the right reasons (which is far from guaranteed in the case of AI systems that can rely on spurious correlations) and if those reasons were combined in a way that makes sense. This can help for both of the problems highlighted here: people under-rely on systems because they cannot easily determine whether the system is correct, and then a cautious approach of trusting your own judgment can make sense (and in fact in the study of Dell'acqua (2022) pays off). Providing explanations that show that the system latches on to the right features in the right way might make it easier to accept that the system is better at the task. Vice versa, good explanations reduce the effort required to check the correctness

of an AI output and can therefore mitigate the issue that human decision-makers stop paying attention during the decision-making.

More generally speaking, while we may not be perfect in explaining why we made a decision ourselves, the resulting explanation is still one that can (and is) judged by others when they in turn have to decide whether or not to agree with us. In other words, the practices of explaining ourselves when stakes are high have probably emerged because they improve these processes. Giving and parsing explanations is cognitively intensive work, and we probably wouldn't be doing it if it wasn't important to our joint success. That will be due in part to the ability to determine whether we agree with a certain claim or not. But, we also provide explanations because they allow for a broader set of corrective measures that go beyond the scope of appropriate reliance as discussed in this subsection. I will therefore consider this additional option for improving outcomes next.

3.2 Correcting Mistaken Decisions

Immediate human oversight is an important part of the way AI system outputs lead to decisions in high-stakes settings, but typically there is a broader set of measures that impact the (long-term) quality of decision-making. Broadly speaking, these can be seen as measures taken to correct mistakes resulting from the initial decision, through accountability and contestability measures. If we only focus on reliability, and accept a system that provides no more than an output and little to no additional information, then there is a risk that correcting mistakes becomes much harder. That, again, leads to worse outcomes despite the fact that the AI system may initially seem to perform better when looking only at its own accuracy.

To start with the clearest way in which this would happen, contestability procedures typically rely on the availability of reasons for a decision. While the notion of contestability is quite broad (Lyons et al., 2021) and can include the reliance on an AI system by the decision-maker, it is here interpreted as the steps that decision subjects can take after receiving a decision they disagree with. Part of this can be algorithmic recourse, where decision subjects are provided with ways in which they can change their situation in order to obtain the desired outcome (and part of the motivation for Wachter et al. (2017) to introduce counterfactual explanations), but it also encompasses cases where a decision is mistaken due to faulty data, missing information, or incorrect reasoning (see also Ploug and Holm (2020) who list data, bias, performance, and decisional role as dimensions for contestability). Many of these dimensions require transparency (e.g. to know what data was used, in order to be able to correct mistakes and spot missing information) but when it comes to objections relating to reasoning there seems to be no good alternative to explainability. One may contest the argumentation behind a decision, after all, without disagreeing on the facts. In order to be able to provide a good counter-argument, we first need to know what lead to the decision in the first place. In these cases, access to data and performance metrics is simply not enough.

It may seem as if even here there is a way out, though: Henin and Le Métayer (2021) present a contestability framework that allows decision subjects to propose alternative decision boundaries (more in line with the decision they want) that can

then be tested against the decision boundary maintained by the system. This system then tries to show that it would be better (overall) to take a different decision. A system like that, to challenge the general functioning of the AI system, could certainly be a valuable addition. However, the precise difficulty posed by opaque AI systems is that we are unsure where their decision boundaries are. As a result, testing alternative norms for decisions would be hard to put into practice unless we can clarify how the alternative differs. If it is merely that we can point to an alternative system that gives us the result we want, then the justification for switching to that alternative remains unclear. There must be something about the contested case that was processed incorrectly, so if we want to ensure that these mistakes are not repeated in the future it is important to know why this alternative system is better.

That brings us to accountability mechanisms, as these are generally intended to ensure that mistakes do not reoccur in the future. When a mistake has been made, the challenge is to find a reason why that mistake happened and the people responsible for fixing it. In cases where AI decision-support plays an important role, the risk is that (for opaque systems) such accountability is difficult to provide. The decision-maker, as discussed earlier, may have simply relied on the system and so the buck is passed on to the AI system. Yet, due to the opacity of this system we may not know why it produced an erroneous output. This issue is well represented in the literature, typically in the context of responsibility gaps (Santoni de Sio & Mecacci, 2021). Although there is a question whether responsibility is really absent in any of these cases (Hindriks & Veluwenkamp, 2023; Königs, 2022), it is nevertheless clear that providing accountability (i.e. explanations or reasons (Binns, 2018)) after the fact can be greatly complicated by the opacity of AI systems.

However, to some extent, the diagnosis of mistakes can be done on the basis of analyzing patterns in the AI outputs. Tesla's misclassification of the full moon as an orange traffic light is easily understood as a case of spurious correlation based on the shape and color of the moon. However, many of the mistakes will be more subtle, as if it was easy to pinpoint why an AI system makes mistakes in edge cases or in what way it incorrectly generalizes, it would also be much more feasible to correct these mistakes. Explanations could help here, as it seems likely that if we have a better idea of why the system fails to work as intended it might allow for more targeted fixes. Perhaps, though, this is to some extent wishful thinking. Targeted fixes are difficult not just because we don't always know why a system makes a mistake, but also because we have few options to intervene in system behavior beyond (re)training it with additional or different data. Mechanistic explainability techniques, such as those pioneered by Anthropic (Templeton et al., 2024), do yield some promises of isolating specific parts of a neural network that cause output behavior. Changing these internal parameters also changes the output behavior of the AI system, although not always in predictable or intended ways as the many interactions between the different parts of a neural network still make it hard to oversee the consequences of these targeted changes. It is, however, a good example of an explainability technique that gets closer to philosophical ideas of explainability (e.g. Buijsman (2022)) while also showing promise for allowing us to fix mistakes directly and improve model robustness (Zhang et al., 2024).

Both contestability and accountability help to improve decision-making procedures over time. They're not directly captured by the technical reliability of the AI system at deployment, but having mechanisms to correct both individual decisions and overall model performance over time likely makes a big difference over time. The final outcomes will be better if we have a good way to correct mistakes. Yet, having that depends on the availability of information. People need to know what data was used, model developers need to monitor the system and actively spot patterns in mistakes in order to fix them, but explanations are also important instrumental tools for people to challenge the reasoning behind a decision and for more targeted, effective fixes.

What all of this shows is that even if we only, or primarily, care about the outcomes of a decision-making procedure then accuracy of the AI system is not enough. AI systems rarely lead to decisions directly, and people need more than just an output and an accuracy score if they are to critically reflect before making a decision. Procedures to correct mistakes likewise depend on the availability of more information than just the final decision. This gives us a new argument in favor of explainability, on the terms of those who have argued against it. Before discussing the consequences of such an argument it makes sense, however, to see what additional reasons for explainability there might be. These tend to be based more on fairness, legitimacy and other aspects typically covered under the heading of procedural justice and therefore I've collectively termed them as such, to make the difference in the line of argumentation explicit.

4 Procedural Arguments for Explainability

So far I have focused on outcome-based arguments relating to a need for explainability. However, we need not agree with the critics of such a requirement that outcomes are the main (or perhaps the only) thing we should care about in high-stakes decision-making. This section therefore looks at procedural arguments, in order to give a proper basis for the proposal on how to handle a requirement for explainability (in Sect. 5). By going through the procedural arguments it is possible to see that even if we accept all these points, we can still find a way forward with AI in high-risk settings. Now, generally speaking, such accounts of whether a procedure is just have been subject to relational critiques, such as those of Meyerson and Mackenzie (2018); Meyerson et al. (2021). In their words: "procedural justice requires more than the use of reliable procedures; that it requires defendants to be offered the opportunity to be heard and present evidence on their own behalf even if this is not necessary to reach the right result; and that perceptions of procedural justice should be taken seriously." (Meyerson and Mackenzie (2018), p.9) These considerations are no less weighty, and at least as far as the perceptions of procedural justice go we find the same results in people's engagement with algorithms (Lee et al., 2019). If we were to follow this relational account in the justification of (algorithmic) decision-making procedures, then it would be necessary to ensure that people have a voice in these procedures and that it is clear that their interests are taken into account during the decision-making.

In a way, many of the proponents of a requirement for explainability have followed this line, though not explicitly so. Vredenburgh (2022) appeals to the right to informed self-advocacy as the basis for a right to explanations. This right incorporates the ability for individuals to present their interests (i.e. to have a voice in the procedure) as well as "forward-looking exercises of agency to navigate systems of rules to achieve one's goals, and backwards-looking exercises of accountability to remedy mistakes or unfairness." (Vredenburgh (2022), p.213) The exercises of agency to achieve ones goals align with the the argument made by Vaassen (2022) that explainability is necessary for people to maintain the ability to shape their lives according to their goals and preferences. Arguably, this too will lead to better outcomes overall, as people will be better-off if they have effective means to shape their lives. Likewise, the ability to hold people to account for mistakes and unfairness was mentioned already in the previous section as one way in which outcomes can be improved beyond the reliability of the AI system. The emphasis, however, is on the ability of people to do this themselves. Even if the outcomes were exactly the same, it matters whether people had the ability to steer these outcomes themselves. Chan (2023) takes a similar relational line in the context of high-stakes medical decision-making. According to him, "Treating patients with effectiveness and respect for their dignity and autonomy requires being able to explain medical diagnosis or treatment recommendation." (Chan (2023), p.287) Explanations are needed in order to maintain autonomy in the face of automated decision-making. For without explanations, the reasoning goes, decision subjects lack the information to anticipate how they should present their case to get to the result they want, and will find it harder to contest decisions after the fact that they disagree with.

This idea, that people should remain in control of their lives even as AI systems are integrated in decision-making, connects to the framing given by Lazar (2024) in terms of explainability as a requirement for democratic legitimacy. He proceeds from the point that AI systems intensify and introduce power relations, and that "power relations must meet standards of procedural legitimacy and proper authority. This is necessary for them to protect and realise democratic values of individual liberty, relational equality, and collective self-determination." (Lazar (2024), p.28) Explanations are needed in order to counter-act the power that AI systems would otherwise hold over decision subjects. Here, too, even if the systems lead to optimal outcomes, that would not suffice to ensure their democratic legitimacy. The lack of self-determination and counter-balance to the power wielded through the use of AI would still be problematic, even if they were used for the common good.

I believe we have good reason to care about these more procedural aspects of (high-stakes) decision-making, in addition to the outcomes that the procedures lead to. Autonomy, self-determination and legitimacy are all important values that should be upheld, even as we integrate AI into our procedures. Often, as discussed in the previous section, following these precepts will also have the benefit of improved (long-term) outcomes, though for most of the authors discussed in this section that is not the main concern. Yet the question is how far these arguments take us. Transparency frameworks presented by Loi et al. (2021) and Buijsman (2024) are also designed to establish the legitimacy of AI systems, and to allow for informed self-advocacy in relation to these systems. They are based on the idea that people need

information on what the goals of the system are, how those have been translated into (socio-)technical specifications and how well the AI system then manages to meet those goals. Importantly, this wouldn't require explanations of the inner workings of the AI system, but only information on things such as the objective function, performance metrics, biases, and effectiveness of human oversight and contestability mechanisms. That could already give decision subjects a good basis for protesting the use of specific AI systems as not in their interests, providing backward-looking accountability and forward-looking agency with respect to the system as a whole. Furthermore, having an idea of what the system outputs, to what purpose, and based on what information, may also help in planning for your own interactions with the relevant decision-making procedure. However, challenging individual decisions would still be difficult without explanations of the underlying reasoning for those decisions. As a result, transparency only provides decisions subjects with means for self-advocacy on a system-wide level, and misses out on the level of individual decisions to be made. Since mistakes can be made on that level too, without translating into wholesale unacceptability of the AI system, this matters. In my view we therefore have two complementary arguments in favor of an explainability requirement. Since we do not currently seem to be able to meet this requirement, the question then is: what implications does this have for our use of AI systems?

5 What if We Require Explanations?

So far I have argued that we have both outcome-based and procedural reasons to ensure that AI systems are explainable. Effective human oversight seems to depend on being able to get explanations, contestability and accountability mechanisms likewise need explanations in order to challenge and correct the reasoning behind a decision. Furthermore, values of autonomy and self-determination point towards an explainability requirement when we look at AI systems at the level of individual decisions. Yet, how strong is this requirement for explanations? Since we cannot offer proper explanations for AI systems yet, a hard requirement would entail that we shouldn't use AI in high-risk settings despite potential improvements in decision-making accuracy. I will argue here that the requirement should not be interpreted that strictly, but that it nevertheless gives us a good reason to avoid implementing AI systems whose output content is closely related to the content of the final decision being made.

To start, there is a clear limitation to the outcome-based argument I've presented: if a particular opaque AI system, despite all the challenges to appropriate reliance and correcting mistakes, still leads to better results then there is no outcome-based argument against using that AI system instead of a less accurate but explainable process. That being said, the range of studies discussed in sect. 3 should make us skeptical that this will be easily achieved in practice when people are asked to either directly follow the AI output or to actively disagree with it. On the other hand, there have been successes with AI implementations where the roles are more complementary and the AI outputs are not as closely linked to high-stakes decisions.

Zhang et al. (2024) is an excellent example of this. Instead of directly predicting the risk of sepsis in a patient, the developed AI system outputs hypotheses to consider along with suggestions for which additional tests to run in order to decrease the uncertainty about the patient's diagnosis. It does not give a direct recommendation on whether or not the patient has sepsis, but instead supports the earlier hypothesis formation and evidence gathering stages of the decision-making process that lead up to the final decision. As a result, the final decision can be explained fully in terms of the performed tests that clinicians understand and are familiar with, without needing to refer to the AI system. The upshot is that we can improve the quality of the final decision, but in a way that limits the risks to over/under-reliance (assuming that clinicians remain skilled at determining which tests are needed, but note that their skills at interpreting these tests in any case remain intact by this setup) and to the contestability of medical decisions. Going one step further, Dembrower et al. (2023) discuss a radiology screening procedure for breast cancer where AI helps identify which people need follow-up tests. Instead of the standard two radiologists doing the screening, they tested a setup where the screening is done by one radiologist and (independently) by one AI system. Whenever one of the two flags potential breast cancer, the patient was then referred to the hospital for further testing, leading to both a big efficiency gain and an improvement in screening sensitivity. This is a high-stakes decision itself, as false positives can be very costly, but not one where the final diagnosis is dependent on the AI system or where the decision of the radiologist to flag a case for follow-up is influenced by it.

Another case of an AI system independently making decisions, though even more integrated into the decision-making procedure, is that described by Zerilli et al. (2019). They discuss a setup at a health insurance company where AI systems do the initial screening of insurance claims. If the system outputs a sufficiently high score, they are automatically paid out (and the company only performs regular checks of these to maintain the accuracy of the system). If the score is below the threshold, the claim is sent to an employee who then evaluates it without any AI assistance, under the framing that this is a complicated case that needs human attention. A sizable portion of these claims is indeed paid out in the end, plus any rejected claims have been carefully examined by an employee. So again, the AI system need not feature in the explanation for why the claim was rejected (the relevant decision for contestability).

What these cases show is that we can have AI systems in high-stakes decision-making procedures contributing substantively to the quality of that procedure, but without causing the problems highlighted in sects. 3 and 4. Reliance on these systems is not an issue, as the actual decisions are made independently by people, without relegating them to correcting the AI system. The decisions that are automated (approving insurance claims in the last example) can then be justified in terms of their reliability, fitting e.g. computational reliabilism but also my broader outcome-based requirement. Contestability and accountability are also maintained: for the sepsis case there is an independently motivated diagnosis, for the insurance claims every rejected claim has a person that can explain their reasoning (and approved claims do not require contestation or accountability) and for the cancer screening there is a radiologist who also looked at every person, as well as one that handles the follow-up appointments. I also believe that these procedures meet the requirements coming

from the procedural arguments for explainability. People maintain the ability to be heard in these cases, as they can e.g. discuss their rejected claim with the person who handled it. People's autonomy is also kept intact: in the insurance claim setup, where self-determination is most directly applicable, there are still clear standards for rejecting a claim that can be communicated to those filing one. Here, I argue, we can both use an opaque AI system and satisfy the critics of opacity in high-stakes decision-making. We can have AI systems aid in high-stakes decision-making without harming our ability to fully explain the (relevant) decisions both among decision-makers and towards decision subjects.

That being said, the arguments for explainability still impose limits to the kinds of AI systems that we can implement. Many of the systems discussed earlier, even if they in isolation are highly accurate at the time of deployment, simply wouldn't meet the bar once implemented within a socio-technical system. The direct link between AI output and final decision greatly limits the ability to critically reflect on and explain the final decision, reducing the quality of the decision and the procedural justification for it. These together give us a clear reason to look for alternative roles of the AI system, but not for a total ban of AI usage in decision-making.

6 Conclusion

To what extent do AI systems have to be explainable in order to use them? I have argued that in some cases they don't have to be explainable at all. However, this is not because their reliability trumps all other considerations. Rather, it is because when suitably decoupled from the final decision, they can be valuable to the decision-making without harming our ability to reflect on and explain the final decision. In these cases we avoid all of the issues raises in this paper: (1) the challenge to appropriately rely on AI systems in the absence of explanations, (2) the difficulty to ensure contestability and accountability of decision-making procedures that correct mistakes when no explanations are available and (3) the difficulty to protect autonomy and self-determination when using opaque AI systems in decision-making. These reasons together give a compelling argument against a range of AI implementations, especially those where the content of the AI output is the same as (or closely related to) the content of the final decision. Having an accurate AI system, then, is not enough. However, that doesn't mean that opaque AI is unfit for use until we can make it sufficiently explainable. We can find ways to benefit from AI systems without making procedures unjust, but have to carefully construct the socio-technical systems that they are a part of if we want to succeed.

Data Availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The author(s) declare that there are no conflict of interest.

References

Beisbart, C. (2021). Opacity thought through: on the intransparency of computer simulations. *Synthese, 199*(3–4), 11643–11666.

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 *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, pp. 248–266, New York, NY, USA. Association for Computing Machinery.

Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology, 31*(4), 543–556.

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines, 32*(3), 563–584. Publisher: Springer.

Buijsman, S. (2024). Transparency for AI systems: A value-based approach. *Ethics and Information Technology, 26*(2), 34.

Candelon, F., Krayer, L., Rajendran, S., & Zuluaga Martinez, D. (2023). How people can create–and destroy–value with generative AI.

Carey, A. N., & Wu, X. (2023). The statistical fairness field guide: Perspectives from social and formal sciences. *AI and Ethics, 3*(1), 1–23.

Caton, S. & Haas, C. (2020). Fairness in Machine Learning: A Survey [cs, stat]. Preprint retrieved from arXiv:2010.04053

Chan, B. (2023). Black-box assisted medical decisions: AI power vs. ethical physician care. *Medicine, Health Care and Philosophy, 26*(3), 285–292.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*, pp. 307–317, College Station TX USA. ACM.

de Jong, S., Paananen, V., Tag, B., & van Berkel, N. (2025). Cognitive forcing for better decision-making: Reducing overreliance on AI systems through partial explanations. *Proceedings of the ACM on Human-Computer Interaction-CSCW, 2025*, 1–30.

Dell'Acqua, F. (2022). Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters.

Dembrower, K., Crippa, A., Colón, E., Eklund, M., Strand, F., ScreenTrustCAD Trial Consortium. (2023). Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet, 5*(10), e703–e711.

Duran, J. M. (2024). Beyond transparency: computational reliabilism as an externalist epistemology of algorithms.

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*(5), 329–335.

Dziugaite, G. K., Ben-David, S., & Roy, D. M. (2020). Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability [cs]. Preprint Retrieved from arXiv:2010.13764

Enqvist, L. (2023). Human oversight in the EU artificial intelligence act: What, when and by whom? *Law, Innovation and Technology, 15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

Hase, P. & Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? [cs]. Preprint retrieved from arXiv:2005.01831

He, G., Buijsman, S., & Gadiraju, U. (2023). How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM Human-Computer Interaction, 7*(CSCW2), 1–29.

He, G., Kuiper, L., & Gadiraju, U. (2023b). Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*, pp. 1–18, New York, NY, USA. Association for Computing Machinery.

Henin, C., & Le Métayer, D. (2021). A framework to contest and justify algorithmic decisions. *AI and Ethics, 1*(4), 463–476.

Hindriks, F., & Veluwenkamp, H. (2023). The risks of autonomous machines: From responsibility gaps to control gaps. *Synthese, 201*(1), 21.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery, 9*(4), Article e1312.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. Publisher: Springer.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, Article 108352.

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology, 33*(3), 487–502.

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology, 24*(3), 36.

Lazar, S. (2024). Legitimacy, authority, and democratic duties of explanation. In D. Sobel & S. Wall (Eds.), *Oxford Studies in Political Philosophy* (1st ed., Vol. 10, pp. 28–56). Oxford: Oxford University Press.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM Human-Computer Interaction, 3*(CSCW), 1–26.

Levin, T. (2021). Tesla's Full Self-Driving tech keeps getting fooled by the moon, billboards, and Burger King signs.

Liu, L. T., Raji, I. D., Zhou, A., Guerdan, L., Hullman, J., Malinsky, D., Wilder, B., Zhang, S., Adam, H., Coston, A., Laufer, B., Nwankwo, E., Zanger-Tishler, M., Ben-Michael, E., Barocas, S., Feller, A., Gerchick, M., Gillis, T., Guha, S., Ho, D., Hu, L., Imai, K., Kapoor, S., Loftus, J., Nabi, R., Narayanan, A., Recht, B., Perdomo, J. C., Salganik, M., Sendak, M., Tolbert, A., Ustun, B., Venkatasubramanian, S., Wang, A., & Wilson, A. (2025). Bridging Prediction and Intervention Problems in Social Systems. Preprint retrieved from arXiv:2507.05216 [cs].

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology, 23*(3), 253–263.

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

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–25.

Meyerson, D., & Mackenzie, C. (2018). Procedural justice and the law. *Philosophy Compass, 13*(12), Article e12548.

Meyerson, D., Mackenzie, C., & MacDermott, T. (2021). *Procedural Justice and Relational Theory: Empirical, Philosophical, and Legal Perspectives*. Taylor & Francis.

Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C., & Faisal, A. A. (2023). Quantifying the impact of AI recommendations with explanations on prescription decision making. *npj Digital Medicine, 6*(1), 206.

Nesvijevskaia, A., Ouillade, S., Guilmin, P., & Zucker, J.-D. (2021). The accuracy versus interpretability trade-off in fraud detection model. *Data & Policy, 3*, Article e12.

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics—A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine, 107*, Article 101901.

Rashid, A. B., Kausik, A. K., Sunny, A. H., & Bappy, M. H. (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International Journal of Intelligent Systems, 2023*(1), 8676366.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495–514.

Robbins, S. (2025). What machines shouldn't do. *AI & SOCIETY, 40*(5), 4093–4104.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*(4), 1057–1084.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23*, pp. 410–422, New York, NY, USA. Association for Computing Machinery.

Shaheen, M. Y. (2021). *Applications of Artificial Intelligence (AI) in Healthcare: A review*. ScienceOpen: ScienceOpen Preprints.

Spatola, N. (2024). The efficiency-accountability tradeoff in AI integration: Effects on human performance and over-reliance. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100099.

Spencer, E.-J., Economou-Zavlanos, N. J., & van Genderen, M. E. (2024). What if we do, but what if we don't? The opportunity cost of artificial intelligence hesitancy in the intensive care unit. Intensive Care Medicine.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N., McDougall, C., MacDiarmid, M., Tamkin, A., Durmus, E., Hume, T., Mosconi, F., Freeman, D., Sumers, T., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet.

Vaassen, B. (2022). AI, opacity, and personal autonomy. *Philosophy & Technology, 35*(4), 88.

Van Der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, Article 103404.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), pp. 1–38. ACM New York, NY, USA.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy, 30*(2), 209–229.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*.

Wang, X. & Yin, M. (2021). Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, pp. 318–328, College Station TX USA. ACM.

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2021). Stop ordering machine learning algorithms by their explainability! An empirical investigation of the tradeoff between performance and explainability. In D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, & M. Mäntymäki (Eds.), *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society* (pp. 245–258). Cham. Springer International Publishing.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector-applications and challenges. *International Journal of Public Administration, 42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In proceedings of the 2019 chi conference on human factors in computing systems, pp. 1–12, Glasgow Scotland Uk. ACM.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines, 29*(4), 555–578.

Zhang, Q., Wang, Y., Cui, J., Pan, X., Lei, Q., Jegelka, S., & Wang, Y. (2024a). Beyond Interpretability: The Gains of Feature Monosemanticity on Model Robustness. Preprint retrieved from arXiv:2410.21331 [cs].

Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L. M., Caterino, J., Zhang, P., & Wang, D. (2024b). Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, pp. 1–18, New York, NY, USA. Association for Computing Machinery.