TU Delft | OPEN Publishing

JHTR
Journal of Human-Technology Relations

WHY SAY WHY?

On the Purposes of AI Explainability

Wessel Reijers wessel.reijers@uni-paderborn.de
Institute of Media Studies, Paderborn University, Germany

Tobias Matzner tobias.matzner@uni-paderborn.de
Institute of Media Studies, Paderborn University, Germany

Suzana Alpsancar suzana.alpsancar@upb.de
Department of Philosophy, Paderborn University, Germany

Article type: Research article

Review process: Double-blind peer review

Topical Collection: Ethics and Normativity of Explainable AI: Explainability as a Social Practice, guest editors: Wessel Reijers, Tobias Matzner, Suzana Alpsancar

This open-access article is published with a Creative Commons CC-BY 4.0 license
https://creativecommons.org/licenses/by/4.0/

DOI: https://doi.org/10.59490/jhtr.2026.4.8233
ISSN: 2773-2266
Submitted: 19 June 2025 Revised: 05 August 2026 Accepted: 07 October 2026 Published: 20 November 2026

How to cite (APA): Reijers, W., Matzner, T., & Alpsancar, S. (2026). Why say why? On the purposes of AI explainability. *Journal of Human-Technology Relations*, *4*, 1-20. https://doi.org/10.59490/jhtr.2026.4.8233

Corresponding author: Wessel Reijers

©2026 The Authors, published by TU Delft OPEN Publishing on behalf of the authors.

Keywords

Explainable AI, Ethics, Purposes, Scope, Value

Abstract

In the wake of rapidly expanding capacities and applications of AI systems that may, to some extent, automate human decision-making, there is a burgeoning debate amongst practitioners, academics, and policymakers concerning the need for explainable AI. While there is a significant consensus around the desirability of explainable AI, there is also a growing body of critique. As explainable AI is supposed to support a major aspect of the responsible innovation of AI, this paper asks: What are the normative and ethical purpose of explainable AI? The paper addresses this question by engaging in a discussion of the normative and ethical purposes of AI explainability. First, it engages in a mapping of different concrete normative purposes that are attributed to AI explainability in the literature. These include, amongst others, diagnosing potential harms, reducing arbitrariness, and enhancing legitimacy. Second, the paper raises a number of limitations faced by explainable AI in realising the purposes it is expected to fulfil, which pertain to the desirability of explainability, its status as an ethical principle, practical obstacles and the rhetorical dangers of explainable AI applications. Third, the paper argues that in order to address these challenges, two over-arching questions should be asked about the extent to which AI explainability can fulfil an ethical purpose: (1) is explainability only merely an instrumental or may it also be a constitutive value, and (2) does explainability pertain to the AI system and its outputs, or to the larger decision-making ecology? Clarifying these questions, the paper argues, is crucial for maintaining explainability as a core ethical principle driving technology development and policy making.

1 INTRODUCTION

The increasing proliferation of AI has been accompanied by concerns over the potential risks and misuse of this new technology. Through initiatives like open letters, policy initiatives, and design interventions, the potential harms of AI are supposed to be reined in. A much-touted principle in the move towards what is widely known as 'responsible AI' is the principle of *explainability*. The principle is not altogether new, as it was already a part of, for instance, data protection regulations (most notably the GDPR) and rule-base expert system development in the early 1990s (Xu et al., 2019). Yet, it responds to a two-faced problem that seems unique to the dominant form of AI, machine learning (ML) on the basis of artificial neural networks. On the one hand, this problem consists of AI being able to augment and even replace human decision-making. For instance, it may decide whether a consumer gets a loan application approved, whether one person rather than another is hired for a job, and whether one route rather than another is followed by a car to prevent an accident. As such, AI affects the control dimension of responsible decision-making: increasingly, the machine and not one or multiple humans are in control of a decision (Matthias, 2004; Sparrow, 2007). This, it must be noted, is not entirely new, as, for instance, financial markets have witnessed the automation of price setting for decades (Mackenzie, 2006). Yet, recent iterations of AI project this new form of automated decision-making (ADM) onto many new areas of (formerly exclusively) human activity. On the other hand, explainability is concerned with the so-called 'black-box' problem (Kamath & Liu, 2021), which is usually linked with the intrinsic properties of neural nets. A particular class of AI models, which are now in focus, operate on many layers of weighted artificial neurons that are configured through a process of supervised and unsupervised training. While other AI models, some of which are now classified as 'Good Old-Fashioned AI' (GOFAI) because they operate on decision trees and allow for some level of transparency in principle (though certainly not always in practice), deep learning models are opaque in principle. This implies that even though humans may understand the inputs (e.g., words in German) and the outputs (e.g., words in English) of a deep learning model, the throughput or the process of 'reasoning' by means of which a model translates between inputs and outputs, cannot be fully understood. This affects the knowledge dimension of responsibility, for responsibility cannot be attributed to a decision that lacked an agent with knowledge about the reasons for the decision as well as its potential consequences.

The hope and expectation of proponents of explainability and explainable AI is that explainability will be the bedrock for responsible AI, addressing the control and knowledge dimensions of AI-driven decisions (Taylor, 2024). The idea is that both a regulatory and technical environment must be constructed that simultaneously makes explainability a legally mandatory feature of a certain class of AI models (especially 'high-risk' models) and offers technical ways to 'open up' the black box, offering both useful and understandable explanations of AI-driven decisions. Driven by research funding and policy initiatives, including the AI Act of the EU, the pursuit of explainable AI is now fuelling multi-disciplinary research and innovation projects, a large range of start-ups, and consulting companies that provide services for responsible AI, and in-house research by large technology corporations like Google and Microsoft. A considerable consensus has been established that explainability should indeed be a new ethical principle that should guide AI development (Floridi et al., 2018). Yet, scholars argue that the current call for explainable AI is both misguided and potentially harmful (Robbins, 2019) or too inflated to be useful (Wasserman-Rozen et al., 2023; Freiesleben & König, 2023; Alpsancar et al., 2025).

This poses an important question, addressed in this paper: what are the normative and ethical purposes of explainable AI? Is it always a good thing to make AI more explainable? When is it good, and when is it not? A comprehensive overview of such purposes and their analysis is thus far lacking. In aiming to answer these questions, the paper relates to what Sven Ove Hansson (2017) has coined the 'ethics of doing ethics of technology'. That is, the pursuit of ethics of

technology – like the pursuit of explainable AI – can itself have negative ethical impacts and should not be treated as self-evident or harmless. For instance, a focus on explainable AI may divert attention and resources from other research into dimensions of responsible AI. It may also generate negative consequences, for instance, when explanations act as a rhetorical foil rather than a genuine means to more responsible decisions. Finally, explainable AI, with the stakes running very high, runs the risk of being subsumed under conflicts of interest, as the motivation for pursuing it will not always be aligned with the original intent of making AI more responsible. With this in mind, an investigation into the ethical purpose of explainable AI does not only hold an academic but also a practical promise, as it may guide more responsible technology development and policymaking.

The paper offers both a mapping of normative purposes of explainable AI[1] and an analysis of these purposes in light of the status of explainability as a central ethical principle. It should be noted from the outset that the thematic of our narrative review is not restricted to 'explainable AI' as it may be narrowly understood, namely as a set of technologies and techniques to make AI systems more explainable and interpretable. While also including this narrow understanding, the paper extends the thematic to what might be coined 'AI explainability,' meaning the normative aim of AI systems to be explainable in particular contexts, which may go beyond the concern with technical solutions (e.g., indicating the need for organizational interventions instead). The paper contributes to at least two strands of research. On the one hand, it contributes to a growing number of reviews of explainable AI (Angelov et al., 2021; Linardatos et al., 2020; Minh et al., 2022), motivated by the desire to capture or map the state of the art of a quickly evolving field. Even though many reviews (partially) address ethics, for instance, in discussing how ethicists define and operationalize explainability, they generally do not address the normative and ethical purposes of explainable AI as such, taking those largely for granted. On the other hand, the paper contributes to a smaller set of literature that explicitly or implicitly takes a stance on the ethical purpose of explainable AI (cf., Jongepier & Keymolen, 2022; Robbins, 2019; Vredenburgh, 2022). While these papers touch upon the theme of this paper, they mostly develop a particular stance on the matter instead of establishing a more general understanding of the debate concerning the normative and ethical purposes of explainable AI.

The paper will proceed as follows.[2] Section one maps different purposes of AI explainability, as these are raised in the literature, organising them according to seven main categories. Section two reflects on these purposes, by emphasising two overarching concerns: (1) whether explainability is merely an instrumental or may it also be a constitutive value, and (2) whether explainability pertains to the AI system and its outputs, or to the larger decision-making ecology. Clarifying these concerns, the paper argues, is crucial for maintaining explainability as a core ethical principle driving technology development and policy making. Section three, then, reflects on a set of important limitations that raise challenges to the possibility of AI explainability in

[1] The mapping exercise has been realised by means of a narrative review (Ferrari, 2015). Rather than identifying, assessing, and synthesizing a well-defined body of literature, a narrative review takes an unstructured approach to extracting the outlines of a general debate that cannot be captured by a single body of literature and offers pathways for future research. Hence, the paper does not aim to be fully systematic and comprehensive, but rather to creatively capture an ongoing debate that is interwoven into different disciplinary discussions of explainable AI. The choice for a narrative review is justified by pointing at (1) the highly dispersed character of discussions about the purpose of explainable AI between different disciplines and relatively isolated academic debates, and (2) the quickly evolving nature of this debate.

[2] This paper was written by the three guest editors of the special issue 'Ethics and Normativity of Explainable AI.' It has followed the regular double-blind review process, with two independent reviewers having reviewed the paper and the guest editors in no way having been involved in the peer review and editorial decision-making process.

realising its normative and ethical purposes. Finally, the paper briefly summarises the findings and reflects on the limitations in the conclusion and discussion section.

2 PURPOSES AND THEIR LIMITATIONS

The discourse on AI explainability is dominated by discussions about the technical realisation of explainable AI and explainable AI as a policy objective. However, these dominant discourses provide insufficient insight into the normative and ethical purposes of AI explainability. The technical literature often treats explainability as a purely epistemic concern, for instance as a means for more meaningful results and detection of adversarial examples (Barredo Arrieta et al., 2019), better hypothesis generation (Došilović et al., 2018); offers a mostly instrumental argument for the purpose of explainable AI, for instance pointing at the use of explainability for generating trust (Miller, 2019); and considers the need for explainability to arise from the mismatch between formal objectives and the dynamism of the real world (Lipton, 2018). As a policy principle, explainability is mediated by a multi-stakeholder dialogue at the intersection of different interests. As such, it suffers from ambiguity (Doshi-Velez & Kim, 2017; Lipton, 2018; Krishnan, 2020), which may lead to lead to 'ethics shopping' (Hickok, 2021, p. 42) and 'ethics washing' (Nannini et al., 2023; Bietti, 2020). Furthermore, as a policy principle, explainability remains largely divorced from the normative context in which it is applied (Mittelstadt, 2019). It lacks a basis for consensus around an ethical principle like beneficence in biomedical ethics that might guide activities in a similar way. Because of these reasons, this paper focuses on a narrower body of literature, which explicitly discusses the normative and ethical purposes of explainable AI, and explainability as a more general principle in the context of AI deployment. This literature offers a myriad of purposes that AI explainability is expected to fulfil.

2.1 PURPOSES OF AI EXPLAINABILITY

This paper identifies at least seven categories of purposes of AI explainability (see Table 1), which may be realized in different contexts. This selection is not meant to be exhaustive but does appeal to being a comprehensive overview of the most important purposes related to AI explainability that lend it its significance as a central ethical principle.

Table 1: Overview of the concrete purposes of explainable AI, and examples of those purposes.

Purposes	Example
1. Diagnosing potential harms	Identifying first-level bias (e.g., racial bias) in a credit scoring algorithm.
2. Enabling responsibility	Allowing medical professionals to take responsibility for adopting or declining an AI recommendation.
3. Providing action-guidance	Providing a citizen with guidance for fulfilling criteria to gain access to a social benefit.
4. Reducing arbitrariness	Reducing first-level bias in a credit scoring algorithm.
5. Supporting deliberative procedures	Providing explanations for an informed consent procedure in medical decision-making.
6. Cultivating dispositions	Empower citizens to question and challenge AI-driven decisions made by government agencies.
7. Enhancing legitimacy	Provide reasoned assent to adoption of AI decision-making processes.

First, explainable AI may have a diagnostic purpose in exposing bias in decision-making involving an AI system. This links to ethics insofar as it constitutes the first step in mitigating harm (for instance, in clinical decision-making) or alleviating discrimination (e.g., assessing decisions made by a credit-scoring AI). Explanations provided by explainable AI may explain 'how different factors were weighed in coming to a final decision' (Zerilli et al., 2019, p. 676). Yet, for this diagnostic purpose to be fulfilled, 'the explanation must not only describe the relationship of output to machine operation ('what happened inside the machine'), but more importantly, its connection to the original rules and facts' (Ződi, 2022, p. 92). For instance, there should be a connection between a medical diagnosis and the diagnostic facts (i.e., measurements) that are understandable to a medical professional. If explainable AI succeeds in this, it provides a basis for identifying potential harms involved in AI decision-making.

Second, explainable AI may have a *responsibility-enabling purpose* (Coeckelbergh, 2020), i.e., enabling an appropriate agent to take responsibility for the decisions made by an AI system. This is at the forefront of the proposal to establish explainability as a principle in AI ethics, as an explanation helps the agent understand how an AI system works (intelligibility) and to take responsibility for its outputs (accountability) (Adams, 2023, p. 616). This purpose of explainable AI can be compared to the reviewability of legal decisions, which is a 'crucial requisite for democratic governance' (Zerilli et al., 2019, p. 663). Explainable AI, ideally, would help those in charge of assigning responsibility (e.g., the court) to establish whether the responsible agent has fulfilled the appropriate legal requirements and applied due diligence.

Third, explainable AI may have an *action-guiding purpose*. This purpose in part constitutes the beneficence of an explainable AI system, assisting people to change their behaviour to gain certain benefits or mitigate unfair disadvantages. O'Hara takes this to be a general aspect of the process of explaining, as its aim 'is often to facilitate future action' (O'Hara, 2020, p. 4). To understand the possibilities for future action, having reasons for action is crucial (Wasserman-Rozen et al., 2023, p. 20). Hence, citizens need explanations about the reasons for particular decisions that affect them if they are to change their behaviour (Vredenburgh, 2022). Explainable AI may be able to provide those reasons if it can effectively explain why a decision was made (limitations concerning this capacity are discussed below).

Fourth, apart from action guidance, the process of explaining, as made possible by explainable AI, has the purpose of *reducing arbitrariness*. Following Lim, 'without knowing how the AI arrives at decisions, the decisions it issues can justifiably be called arbitrary' (Lim, 2021, p. 288). As AI systems may perpetuate or introduce new biases in a decision-making process, explaining the decisions made can help identify and mitigate first-level arbitrariness or bias, which is an uneven application of standards (Franke, 2022). Arguably, AI systems only have 'extrinsic' and no 'intrinsic' biases (Zerilli et al., 2019), which would make it relatively easier to reduce arbitrariness in AI decision-making as compared to human decision-making. And yet, arbitrariness in the sense of discretion in decision-making may at times also be beneficial, as it allows decision-makers to react to the specifics of a particular situation, much like street-level bureaucrats sometimes must use their discretionary powers to decide on hard cases (Vredenburgh, 2023).

Fifth, explainable AI may support or enhance *deliberative procedures*. This purpose relates to action-guidance, but rather than enabling recipients of a decision to follow existing rules to change future outcomes (e.g., improve one's credit score), it is about deliberating on outcomes and rules, for instance, in challenging them. Agency may be promoted *procedurally*, for example, by offering ways to enable due process in cases where an AI system decides against the interests of a recipient. Explainable AI may offer procedural pathways for people to 'meaningfully participate' in the decision-making process (Colaner, 2022). This procedural concern comes, amongst others, to the fore concerning informed consent in medical decision-making. In this context, it is crucial that the autonomy and agency of patients are respected, and

this can only be effectively done if they can give their informed consent for an AI-driven medical intervention, which would require proper explanation at different levels (Ursin et al., 2023). Crucially, this purpose of explainable AI is only relevant in the context of practices that affect people's 'core deliberative agency' (Jongepier & Keymolen, 2022, p. 7). Whenever people face a question of responsibility (e.g., performing a medical intervention on a patient) in which substantive epistemic goods are at stake (e.g., questions about health), explainable AI may be required to enhance people's core deliberative agency or 'informed self-advocacy' (Vredenburgh, 2022, p. 4).

Sixth, explainable AI may also have the purpose to enhance an agent's *moral disposition*. That is, it may increase the agent's capability to normatively understand and act on a decision. Aligning more with a virtue ethics perspective (Hindocha & Badea, 2022), this perspective sees the purpose of AI to support and promote a form of practical wisdom or *phronesis* (Karlan & Allen, 2024). As Vredenburgh argues, AI-driven decision-making may take place in the context of street-level bureaucracy (e.g., deciding on a benefits application), and this type of decision-making requires particular moral dispositions (Vredenburgh, 2023) or beneficial emotional states like empathy (Ranchordas, 2021). Explainable AI, then, may be developed with the purpose of cultivating the right types of moral dispositions that agents (both decision-makers and decision-recipients) may need to exercise their agency. For instance, explainable AI may empower citizens to question the decisions made by government agencies, enhancing their capability to challenge those decisions.

Seventh, explainable AI may improve the *legitimacy* of a decision-making system as a whole, which is also often framed as 'trust' or 'trustworthiness' (Baron, 2025). This may be understood as providing a sense of basic reliability, leading to a thin sense of trust. It is argued that explainable AI seems to excel at 'strengthening the decision-making system's authority' (Wasserman-Rozen et al., 2023, p. 3). As O'Hara concedes, explanations provided by explainable AI may reassure stakeholders 'that a decision was made conscientiously, carefully and according to accepted methodologies' (O'Hara, 2020, p. 6), which would raise trust in an AI system. A higher bar would be for AI explainability to also contribute to the *legitimacy* of a decision-making system, which would lead to people having legitimate reasons to follow its decisions (Wasserman-Rozen et al., 2023). This relates to the question of procedural agency, as the legitimacy of a decision-making system is undermined if citizens do not possess reasons helping them to intentionally adjust their behaviour (Vredenburgh, 2022, p. 4). Over and above procedures that individuals may follow to increase their agency and hence the legitimacy of a decision-making system, explainable AI may become part of what is coined as 'democratizing AI' (Noorman & Swierstra, 2023). This would require AI-driven decision-making to be not only explainable but also contestable, which would allow decision-makers and recipients of decisions to participate in the shaping of the decision-making ecology.

In short, explainable AI may ideally serve many purposes, including diagnosing harms, enabling agents to take responsibility, providing normative guidance for actions, reducing arbitrariness in decision-making, enhancing the agency of decision-makers and recipients of decisions by strengthening procedures and moral dispositions, and enhancing the legitimacy of the decision-making systems. Yet, the realisation of these purposes through the development of explainable is faced with a number of challenging limitations, which the paper turns to now.

2.2 LIMITATIONS OF AI EXPLAINABILITY

Given the previous discussion of purposes, it is not surprising that the notion of AI explainability is attached to some significant hopes and expectations. If it lives up to its purposes, it could indeed be a cornerstone of responsible innovation, as it would enable responsibility through intelligibility and control, but also an increase in human agency and more robust and legitimate decision-making ecologies. However, despite there being an apparent consensus on the

desirability of AI explainability (Babic & Cohen, 2023), there is a surge in criticisms of explainable AI that point to its current and potentially inherent limitations. Many limitations of AI explainability are increasingly being identified and discussed in the literature. Discussions surrounding these limitations focus on the purpose of AI from different angles, which we will survey. First, the most stringent critique focuses on the desirability of explainability and explainable AI, whether it is at all or, in all instances, a good thing. Second, explainability is questioned insofar as it is and should be an ethical principle worthy of being pursued in policymaking and technology development. Third, while acknowledging the desirability of explainability, several scholars critique its practical feasibility. Fourth, there is a general worry that explainability merely constitutes a rhetorical device, not an ethical response.

To start, scholars have questioned the very *desirability* of explainability in the context of AI. One of the fiercest critics is Robbins, who argues that imposing explainability for AI decision-making is not only misguided but also potentially harmful (Robbins, 2019). He argues, first, that explainability should be applied to a decision and not to the decision-maker (e.g., an AI system), and that not all decisions are equal – that many do not require an explanation. This point resonates with other criticisms (Jongepier & Keymolen, 2022; Zerilli et al., 2019) that argue against having an asymmetry or double standard concerning decision-makers (humans or machines). Second, and more importantly, if we want to use a machine learning application for a particular problem, spelling out the considerations that led to the decision would defeat the application's purpose, as it functions well precisely because the considerations cannot be made explicit and translated into a computational decision-tree. The potential harms caused by AI, Robbins argues, are not due to a lack of explainability but a lack of efficacy. Another course of reasoning argues against the value of transparency and reduced transparency in all cases. Transparency does not necessarily lead to understanding (i.e., not all explanations foster understanding), which is known as the fallacy of transparency (O'Hara, 2020). Although explainability is not the same as transparency, it can lead to it, which means that if transparency is undesirable, explainability may be similarly so. What is more, in certain decision-making contexts, explanations leading to transparency are harmful or even forbidden. Wasserman-Rozen and colleagues argue, for instance, that in the case of decision-making by juries in the context of criminal cases, explaining this decision is explicitly forbidden (Wasserman-Rozen et al., 2023, p. 28). In a similar vein, Vredenburgh argues that the lack of discretion in the decision-making of AI systems, which may be exacerbated through explainable AI that explains which rules were followed, is not always beneficial as it may generate organizational pathologies due to a lack of certain moral dispositions (Vredenburgh, 2023). The general point of criticism emanating from these evaluations is that explainability in AI is not always desirable and may even be counter-productive or harmful.

A second set of criticisms tackles the *practical limitations* of current approaches to explainable AI. These criticisms are largely based on a comparison between human and AI decision-making (including explainable AI decision-making), arguing that the latter lacks aspects of what the former provides. Ződi (2022) enumerates these differences, arguing that machines lack the discretionary power humans have, the capacity to interpret, the capacity to narrate facts, forward-looking capacities, and the perception of (human) value. Where these aspects are crucial to human decision-making, like in a legal context, explainable AI should not be used. Specifically, current explainable AI approaches are limited in two regards. First, they only offer post-hoc explanations. Such explanations can only offer approximations of the decision-making process (Hatherley et al., 2026, p. 170). Furthermore, they offer explanations that are not unique (do not capture the singularity of an AI-produced decision), and not sincere (because they may be inconsistent between cases) (Babic & Cohen, 2023). Second, unlike humans, AI systems have no direct feedback mechanism between the process of explanation and the process of learning. In other words, they lack regulative reason (Peters, 2023). This, however, is essential for the way explanation in human decision-making contributes to its trustworthiness,

as it produces lasting commitments. As such, this argument opposes the insistence of other critics opposing a distinction between explainability for humans and machines. On this basis, some scholars argue that the wish to pursue accuracy, inadequately balanced through a level of explainability, fails to justify a rejection of interpretable AI (i.e., AI based on transparent decision trees) (Hatherley et al., 2026). Overall, these criticisms of explainable AI argue that it cannot live up to some crucial aspects of human decision-making, rendering it unsuitable for achieving the standards of responsibility and transparency it is supposed to live up to.

Third, explainability is criticised for being a *questionable ethical principle* to guide AI development and policymaking. The main concern here, is that explainability occupies a position in between an epistemic and an ethical demand, which is acknowledged by the proponents of explainable AI (Floridi & Cowls, 2019). As Russo and colleagues argue, ethics and epistemology remain largely disconnected in the discourse on explainable AI (Russo et al., 2023). It is one thing to provide an explanation or to explain, it is another thing to do so in a way that answers an ethical concern. This resonates with a point made by Robbins, who argues that 'ethical value is to be contrasted to the epistemic value explicable AI might provide' (Robbins, 2019, p. 501). As evaluative outputs of AI systems cannot be verified in the proper sense, according to Robbins, they should not be used when dealing with questions of value (Robbins, 2025). As such, explainability shares an ethical ambiguity with a principle like security. Security, it may be argued, is not by itself an ethical principle as it can be beneficial (e.g., securing one's personal information) and harmful (e.g., securing a prison with political dissidents). Explainability, similarly, is not necessarily beneficial, as reasoned above, which makes it different from principles like beneficence and justice. Some authors criticize explainability on the basis of its incompatibility with other substantive ethical principles. Even though explainability may be epistemically necessary for respecting ethical principles like justice, it should not be taken as an ethical principle in and of itself (Cortese et al., 2023). Even if taken as an ethical principle, explainability may suffer from being too vague, conceptually unclear, and capable of conflicting with other principles (Ursin et al., 2023). This leads to a related criticism, which emphasizes that explainability, instead of a principle, may be a factor in a cost-benefit consideration (Kempt et al., 2022). It may be preferable to sacrifice explainability in certain cases if other ethical concerns outweigh it.

Fourth, critics have raised the suspicion that explainability is a *rhetorical device* rather than a genuine means for gaining understanding – a 'fool's gold' (Babic & Cohen, 2023) and 'ethical fig leaf' (Starke et al., 2023). Post-hoc explanations offered by explainable AI, it may be argued, only offer after-the-fact rationalizations of decisions that do not explain the decision-making process itself and thus offer a 'kind of moral sedative' (Babic & Cohen, 2023, p. 891). The drive to rhetorical fabulation may be inherent to explainability, as the demand for transparency that comes with it can force experts to act only in ways with explainable AI that afford public justification (Cortese et al., 2023). Experts themselves, in an empirical study, express the fear that explainable AI may function as an ethical fig leaf. It may be used for 'covering methodological shortfalls by providing end-users with a false sense of understanding' (Starke et al., 2023, p. 311). At the back of these concerns may lie a more general issue concerning the distinction between explaining and justifying, or the gap between first-level and second-level bias (Franke, 2022; Hildebrandt, 2020). That is, while explainable AI may provide explanations of first-level bias, it cannot justify the existence of second-level bias – yet, it can provide illegitimate rationalizations for such biases, generating a greater sense of trust and authority (Wasserman-Rozen et al., 2023). In turn, the incapacity of explainable AI to provide justifications may, in time, undermine the legitimacy of AI-driven decision-making (Babic & Cohen, 2023). This has already caused 'some authorities [to] have begun to contest the reality of the accuracy-interpretability trade-off in AI and ML' (Hatherley et al., 2026, p. 172). In line with the limitations concerning legitimacy, Baron (2025) argues that explainability may be needed for

certain strong forms of trust, but that AI explainability is not an appropriate response to such needs.

3 SCOPE AND VALUE OF AI EXPLAINABILITY

Thus far, the paper has unearthed seven categories of purposes that AI explainability is supposed to realise, while putting forward striking reasons why there are (strong) limitations in the way of this realisation. To understand how we might deal with this tension, we propose to consider two over-arching concerns that we argue need to be considered to progress the work on AI explainability. The first concern has to do with its scope, whether it pertains to ideally automated explanation generation by an AI system or to the decision-making ecology writ large. The second, related, concern has to do with its ethical value: whether AI explainability has a constitutive value, understood as contributing to the realisation of ethical purposes that could otherwise not be realised, or a mere instrumental value, understood as a factor that may contribute to the realisation of certain ethical purposes but not necessarily so.

3.1 SCOPING AI EXPLAINABILITY

The question concerning the purpose of explainable AI first confronts us with the dilemma of how *purpose* is meant in the first place. In other words, it leads to questions of what it is that has a purpose, how purpose is understood, and which relations purposes have with one another. These are not yet questions about the concrete uses of explainable AI, for instance, whether it may assist a doctor in healing a patient, but rather overarching questions about its purpose that help us draw basic distinctions. So, first, when we talk about explainable AI, what do we refer to when considering its purpose? This raises the question of scope, which may be narrow or broad. Narrowly conceived, the purpose of explainable AI would be limited to an AI system and the technical models that explain its behaviours. Some authors argue that only this narrow understanding of explainable AI is useful and that the explainability criterion should be understood as a 'formal-procedural criterion' (Ződi, 2022, p. 91) exclusively reserved for machine decisions or outputs. Ződi compares explainability in the context of algorithmic decisions with juridical decisions and argues that these are very distinct. Unlike human juridical decision-making, algorithmic decisions could (arguably) not have been made otherwise, due to a lack of meaningful alternatives, the narrative dimension of juridical decisions, and a genuine forward-looking perspective (Ződi, 2022, pp. 83–84). For this reason, enlarging the scope of explainability beyond formally probing the outputs of AI systems makes the concept conflated and risks confusion with other important normative principles like accountability and transparency. Others, however, have argued against the uniqueness of AI decision-making and have pointed to the alleged existence of a 'double standard,' which would subject AI-driven decisions to an unduly high standard compared to 'flawed' human decision-making (Zerilli et al., 2019). Zerilli and colleagues argue that there is no inherent reason for *prima facie* subjecting machines to an explainability requirement that is different from any explainability requirement applied to human decision-making. Similarly, though on different grounds, scholars have argued against the existence of an asymmetry between human and non-human decision-makers (Jongepier & Keymolen, 2022). The decision-making process must be considered in its own right according to this line of reasoning, without having to distinguish between the sites of origin of a decision – whether this was a human or a machine.

This insight opens another perspective on broadly conceiving the purpose of explainable AI. If the purpose of explainable AI is not tied solely to scrutinizing the outputs of an AI system but rather to the process of decision-making as a whole, explainability is no longer a function of just the discreet artifact but of a larger ecology. As O'Hara argues, 'if the data subject is to be able to contest a decision, he or she needs to understand the decision-making as a whole, not simply the AI component;' and 'the output of the AI is not the decision of the social system' (O'Hara,

2020, p. 5). In this light, AI systems are never to be considered fully autonomous in the strict sense of the word (i.e., fully independent of human decision-making). Rather, they have a degree of autonomy in a larger ecology of human-machine decision-making. Considered in this larger context, AI explainability may also realise a more general purpose, for instance a holistic understanding of health in the application of medical AI (Herzog, 2022). How should we consider purpose from this broader, ecological perspective? Philosophers of technology (e.g., Ihde, 1979), inspired by thinkers like Aristotle and Heidegger, have considered the purposes of artifacts as being two-sided. In a well-known example of the use of a tool, a hammer can have the simple purpose of hitting a nail into the wall (hammering *in order to* get the nail into the wall). Yet, the artifact in use always refers beyond itself, towards a network of purposes, for instance, getting the nail into the wall *in order to* support the roof. Eventually, this network of purposes terminates in final ends, or an end (*telos*) that completes an activity (Heidegger, 1977, p. 8). Eventually, hammering happens *for the sake of* providing shelter for human beings. Similarly, an explanation provided by an explainable AI system will derive its purpose not only from its direct use (e.g., avoiding a false positive in deciding on a medical treatment) but from its relation to other purposes, which in turn relate to a possible 'final end' that completes the activity, like promoting health. It is only from the perspective of the larger ecosystem, in which these final ends are embedded, that we may speak of the purposes of explainable AI.

The diversity and interrelation of purposes of explainable AI seem to relate to the diversity of practices of explanation and what may be coined as levels of abstraction. To illustrate, while not considering explainability but accountability, Wieringa argues that 'the interests and foci of the respective accounts deviate across the different kinds of accountability practices' (Wieringa, 2023, p. 19). In the wake of the so-called childcare benefits scandal in the Netherlands, in which algorithmic decision-making contributed to thousands of citizens being illegitimately labelled as frauds by the tax authority (although the system in question probably did not involve machine learning), different 'accountability gaps' emerged (i.e., a technology-induced gap between a given practice and the extent to which accountability standards can be successfully applied to this practice). For instance, the court, in ruling on the legitimacy of an algorithmic system similar to the one involved in the childcare benefits scandal, tied the question of accountability to the technical functionality and deployment of the system, while the media focused on the broader decision-making process around the system. Hence, the purposes of accountability practices, similar to explainability practices, are relative to the different stakeholders involved, an insight that is widely shared in explainable AI literature (Gyevnar et al., 2023, p. 3). A particular stakeholder may relate to a distinct purpose of explainable AI, addressing a distinct 'explainability gap' or 'transparency gap' (Gyevnar et al., 2023). It may be argued that different explainability practices are required for different levels of abstraction in a decision-making system. Scherer makes an argument of this kind when discussing the different levels of explainability involved in explaining a juridical decision. One may explain these by describing (1) the psychological processes of a judge, (2) the procedures that ought to be followed in arriving at a decision, and (3) the standards for appraising a juridical decision, whether it is a good or a faulty decision (Scherer, 2019, p. 29). Similarly, an explainable AI solution may have the purposes of explicating the throughput process of an AI system, assessing the decision-making process according to auditing requirements, and indicating whether an AI system made a good or bad decision and why this is so. Arguably, state-of-the-art explainable AI solutions may be relatively well-suited to address the lower level of abstraction but are not (yet) capable of addressing the higher one.

Reconsidering the scope of AI explainability, we argue, helps addressing the first two limitations, concerning its desirability and the role of practical limitations. To a large extent, the argument against its desirability is premised on a narrow conception of explainable AI, targeting the AI system and explainable AI modules rather than the decision-making system and process as a whole. One way out of the impasse would be to drop explainability as a guiding notion for the

design and policymaking around AI (Robbins, 2019). After all, explainable AI as an entity shares problems of bias and opacity with human decision-makers, and the solution to these problems does not lie in making decisions more explainable or transparent. Yet, as this argument relies on the comparison between humans and machines and is thus 'atomistic' in considering individual decision-makers only, it misses the larger institutional picture (Maclure, 2021; de Bruijn et al., 2022). Instead, we may confirm that explainability is still desirable as an aspect of public reason, but to respect this principle, we need to confront the larger institutional *context* (Russo et al., 2023). Institutions, after all, are capable of addressing and mitigating bias and opacity. This is not just a call for having a 'human in the loop,' but providing both procedural safeguards (Maclure, 2021; Vredenburgh, 2022) and cultivating professional virtues (Vredenburgh, 2023) to confront the risks of AI opacity, thereby avoiding organizational pathologies. Furthermore, in considering the types of procedures and virtues that are appropriate, a more critical approach towards explainability is required: one that does not always consider it as a beneficial thing but also as a potentially problematic aspect of decision-making processes (Wasserman-Rozen et al., 2023). In other words, explainability should not be indiscriminately required or applied, but only in those cases where it is beneficial rather than potentially harmful.

Reconsidering the scope of AI explainability also to some extent addresses the practical limitations. Most of the discussion around these limitations is focused on a narrow understanding of explainable AI, i.e., the limitations of technical solutions to address problems of explainability. Remaining at this technical level, a way to overcome this limitation could be to opt for interpretable AI (or GOFAI) instead of ML AI. Yet even in pursuing this course, it should be considered that interpretable AI does not necessarily bring about transparency in any given context. From a legal and political perspective, there will be cases where interpretable AI is preferred because it provides transparency and a clear link between policy choices and algorithmic design (Babic & Cohen, 2023). Similarly, in medical settings, professionals may be justified in opting for interpretable AI as it allows them to support high-risk decisions (Hatherley et al., 2026). Although this approach circumvents many of the issues, it will not always be appropriate (Robbins, 2019). What will otherwise help to overcome some of the practical limitations is to include a feedback mechanism in explainable AI decision-making. This, similar to human decision-making, will help solidify 'commitments' (Peters, 2023). Even if an AI system by itself lacks the regulative reason needed for such commitments, the larger ecosystem of decision-making arguably does not lack it. That is, humans, in collaboration with explainable AI, are able to engage in a learning process akin to a feedback mechanism that produces commitments. As Russo et al. (Russo et al., 2023) argue, a holistic view of the explanatory process in an institutional context enables linking the epistemic and ethical aspects of explainability. Hence, the challenge is to make sure that the process of explaining does not simply remain a post-hoc rationalization that does not itself influence the decision-making process. Rather, the process of explaining should inform, reinforce, or challenge and contest the decision-making process. In this sense, not only explainability but also (democratic) contestability by design may be needed (Noorman & Swierstra, 2023).

3.2 AI EXPLAINABILITY AS HAVING CONSTITUTIVE VALUE

From here, we get to a second concern about the purpose of explainable AI, namely, is it only instrumentally or also constitutively valuable (Colaner, 2022)? We understand 'instrumental' here as being instrumental to the realisation of another value that could, however, be realised through other means, and 'constitutive' as explainability constituting the realisation of a particular value (i.e., that could not otherwise we realised). A parallel would be the distinction between regulative rules and constitutive rules in speech act theory.[3] The instrumental value of

[3] According to Searle (2006), regulative rules guide behavior and can as such be substituted. Constitutive rules, however, are constitutive to the practice they regulate, i.e., without them, the practice as such would not

explainable AI is relatively straightforward to grasp. For instance, an explainable AI system may be used to show which data points contributed most to a decision concerning a negative credit score, which led to a loan application being rejected. It may turn out that the data points in question are bad predictors, for instance, the number of pets a person has, and that the model was trained on spurious correlations. In this case, the value of explainable AI is one of debugging and fine-tuning an AI model, instrumentally serving its accuracy and effectiveness. Furthermore, it may be that the data points reveal a hidden, problematic bias in the model, for instance, if postal codes were important predictors for the outcomes that correlate with racial groups. In such a case, explainable AI instrumentally serves the value of fairness. Importantly, as long as explainable AI has an instrumental value only, it automatically triggers questions of opportunity costs, i.e., questions concerning the value of foregone alternatives. For instance, as Kempt and colleagues argue, 'the need for at least some quality medical care might trump concern about absolute explainability, which might be considered by some as an unnecessary luxury' (Kempt et al., 2022, p. 8). In other words, explainable AI, even if it is useful, may not constitute the best alternative in light of the higher-order value it is supposed to serve.

It may, however, be that explainable AI is not only instrumentally but also constitutively valuable. Reconsidering the issue of bias, Franke (2022) offers a valuable perspective to grapple with this distinction. Taking Nozick's conception of rationality as a starting point, Franke argues for the distinction between 'first-level bias,' the uneven application of existing standards, and 'second-level bias,' which concerns an explanation of why these rather than other standards were chosen (Franke, 2022, p. 21). Systems that make automated decisions, so Franke claims, are able to check on the existence of first-level bias, for which explainable AI may be instrumentally used, but cannot check for second-level bias. Yet, it is in tackling second-level bias that explainable AI may be intrinsically valuable. Most authors who provide arguments in favour of this view do not directly discuss the value of explainable AI but rather, more generally, the value of explainability, either as a right (Vredenburgh, 2022) or as a principle (Floridi & Cowls, 2019). Explainability, it is argued, is not only instrumentally valuable to realize other values (e.g., fairness by tackling first-level bias), but it is constitutively valuable to human decision-making, offering a basis for what might be called prudential judgment or, in Aristotelian vocabulary, phronesis or practical wisdom (Karlan & Allen, 2024).

At least three pathways are traversed in the literature to argue for the constitutive value of explainability. The first considers this idea in light of a *right to explanation*. Vredenburgh argues that citizens should enjoy such a right, as explanations are necessary to protect a vital interest, namely that of self-advocacy (Vredenburgh, 2022, p. 2). In a similar vein, Jongepier and Keymolen argue that a right to explanation should apply to situations in which our 'core deliberative agency' is at stake (Jongepier & Keymolen, 2022, p. 9). Notably, the value of explainability is not tied to explainable AI under these accounts. Rather, explainability is a right under any 'hierarchical, non-voluntary' institutional decision-making that risks being opaque (Vredenburgh, 2022, p. 4), with AI being an augmenting rather than a transformative factor. A second pathway focuses more explicitly on the intrinsic value of explainable AI and considers explainability as a necessary *principle in AI ethics* and policymaking. As an ethical principle, explainability confers the requirements of intelligibility (an epistemic component) and accountability (an ethical component) on AI systems (Floridi & Cowls, 2019). As such, explainability is not only instrumental in achieving other core principles in AI ethics as derived from biomedical ethics such as beneficence and justice but stands on its own as an ethical principle that should guide AI-related development and policymaking. Since this view has been influential in shaping AI policy, e.g., the EU AI Act, authors have discussed whether explainability should indeed be a stand-alone principle (Adams, 2023) or whether it is already implied in

retain its meaning. Similarly, explainability may be constitutive to particular values like deliberative agency, which without it would lose their meaning.

respect for other substantive principles (Cortese et al., 2023), merely having an epistemic role. This debate reveals an inherent ambiguity at the core of the notion of explainable AI, as it has, similar to prudential judgment, an irreducible epistemic and ethical component (Russo et al., 2023). As such, it relies on what may be considered an 'ethics of argumentation' (Ricoeur, 1992, p. 287). The third pathway, which is sceptical of the promise of explainable AI (see Section 3.1) considers explainability as an *ideal of good governance*. The intrinsic value of explainable AI, from this perspective, should be considered in analogy with the purpose of other pivotal ideals of good governance, like the rule of law (Greenstein, 2022). As Wasserman-Rozen and colleagues argue, reason-giving, which constitutes a basic understanding of explainability, is essential to the functioning of a legal system as it acknowledges citizens' autonomy and offers them the possibility of due process. Hence, explainability serves the non-instrumental purpose of 'securing better and more just decisions' (Wasserman-Rozen et al., 2023, p. 18). This understanding of the constitutive value of explainability leaves aside, for the moment, whether explainable AI can live up to this ideal.

Considering the distinction between instrumental and constitutive value elucidates the role of explainability as a guiding ethical principle. We may follow Adams (2023) in defending explainability on the basis of it being a useful mid-level principle that is not universally but pragmatically applicable. Still, it faces some important challenges. The first concerns the challenge of opportunity cost. Insofar as explainability has an instrumental value, it may be traded off against alternatives. Although the common argument concerns the trade-off between accuracy and explainability, this trade-off may be misguided as accuracy is not necessarily the principal concern of doctors in making medical decisions (Hatherley et al., 2026). What this leads to is a demand for a more systematic way to consider the opportunity cost of imposing explainability, rather than imposing it as a general rule and not taking accuracy as the only trade-off. Promoting a principle like fairness, in certain cases, may be served by downplaying the need for explainability (Kempt et al., 2022). The second concerns the challenge of the gap between epistemology and ethics. Insofar as explainability is intrinsically valuable, it resonates with an ethics of argumentation, both promoting an ethical requirement for discourse and constituting a basis for practical wisdom or prudential judgment. As such, it may promote deliberative agency (Jongepier & Keymolen, 2022) and self-advocacy (Vredenburgh, 2022), which seem constitutive rather than instrumentally served by explainability. To achieve this, as Russo et al. (2023) argue, explainable AI should not remain limited to epistemic concerns, as is the case in much of the literature, but bridge the gap between epistemology and ethics. This may be done by broadening the focus from AI systems to the 'whole process of design, implementation, and use of AI systems' (Russo et al., 2023, p. 1600). Moreover, it would require a democratization process (Noorman & Swierstra, 2023) with a role for experts (potentially free experts for citizens) (Vredenburgh, 2022) and non-experts to assess the decision-making process (Russo et al., 2023).

Relatedly, this let us confront the suspicion of explainable AI being merely a rhetorical device. We may follow Aristotle (ca. 350 B.C.E./2006) in formulating an answer to this problem, insofar as it can be admitted that explainable AI indeed offers means to realize 'the art of rhetoric.' That is, an explainable AI application may be concerned with 'discerning the possible means of persuasion in each particular case' (*Rhetoric* 1.2 1355b26–27). Not all forms of rhetorical persuasion amount to sophistry, as some are in fact crucial for maintaining the integrity of important institutions like the court and parliament. Yet, two important points need to be observed. First, rhetoric knows different modes or genres, and particular genres are appropriate to certain social and institutional contexts. This means that the rhetorical capacities of explainable AI may need to be checked or adjusted depending on the context in which they are applied, for instance, to avoid that an expectation of transparency generates adverse effects (Cortese et al., 2023). This may require specific legal requirements, which, amongst others, 'should thus clarify how the assessment of explainable AI tools is carried out' (Gyevnar et al.,

2023, p. 6). Second, explainable AI will need to provide what Aristotle termed 'proofs,' which may support explanations but also justifications. As Franke argues, this means that explainable AI, viewed from the perspective of a decision-making ecosystem, is able to explain second-level bias (Franke, 2022). This may be done at the level of architecture by articulating the decision-making that led to the integration of certain standards and offering reasons for them. For instance, an explainable AI system would not only explain which factors led to a rejection of an application for a social benefit, but also what higher-level rules govern the process of assessing social benefit applications (Vredenburgh, 2022). It may also be done at the level of organization, for instance, by offering free access to experts who may help citizens contest AI-driven decisions (Vredenburgh, 2022) or facilitating diversity in design situations, which would foster critical thinking concerning methods and standards implied in the design of an AI system (Franke, 2022).

4 CONCLUSION

This paper has considered the normative and ethical purpose of AI explainability. As such, it has tried to tackle the 'why' of explainability, thereby echoing Hansson in doing ethics of AI ethics. First, the paper mapped seven distinct purposes of AI explainability: (1) diagnosing potential harms, (2) enabling responsibility, (3) providing action-guidance, (4) reducing arbitrariness, (5) supporting deliberative procedures, (6) cultivating dispositions, and (7) enhancing legitimacy. Second, the paper discussed the extent to which these purposes may in fact be attained, by considering limitations of AI explainability as it is currently pursued. These limitations have to do with AI explainability not in all cases being desirable, with practical limitations of technical approaches to AI explainability, the questionable nature of explainability as an ethical principle, and the danger of AI explainability becoming a rhetorical foil rather than a genuine solution. To address these limitations, the paper argued that the scope and value of AI explainability need to be reconsidered. Concerning its scope, AI explainability needs to be responsive to the broader decision-making ecosystem rather than merely to the explainable AI system. Concerning its value, explainability ought to be a constitutive rather than merely an instrumental value. Future research should consider these two requirements in the organizational context wherein AI explainability is realised, in which it supports particular discursive practices. This research should investigate what type of normative standards could govern these practices, i.e., how we could distinguish between 'good' and 'bad' discursive practices as realised through AI explainability.

Data Access Statement

N/A

Contributor Statement

Wessel Reijers: Conceptualisation, investigation, writing – original draft

Tobias Matzner: Writing - review and editing

Suzana Alpsancar: Writing - review and editing

Use of AI

During the preparation of this work, the author(s) used no generative AI tools. All writing has been done by hand.

Funding Statement

This work was funded by the German Research Foundation (DFG) under project number 466178508.

Acknowledgments

We would like to thank the journal editors and the anonymous reviewers for their support in improving the paper.

Conflict of Interest Statement

There is no conflict of interest.

References

Adams, J. (2023). Defending explicability as a principle for the ethics of artificial intelligence in medicine. *Medicine, Health Care and Philosophy*, *26*(4), 615–623. https://doi.org/10.1007/s11019-023-10175-7

Alpsancar, S., Buhl, H. M., Matzner, T., & Scharlau, I. (2025). Explanation needs and ethical demands: Unpacking the instrumental value of XAI. *AI and Ethics, 5*(3), 3015-3033. https://doi.org/10.1007/s43681-024-00622-3

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), 1–13. https://doi.org/10.1002/widm.1424

Aristotle. (2006). *The "art" of rhetoric* (J. H. Freese, Trans.). Harvard University Press. (Original work published ca. 350 B.C.E)

Babic, B., & Cohen, I. G. (2023). The Algorithmic Explainability "Bait and Switch". *Minnesota Law Review*, *108*, 857–909.

Barredo Arrieta, A., Diaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baron, S. (2025). Trust, explainability and AI. *Philosophy & Technology*, *38*(1), 4. https://doi.org/10.1007/s13347-024-00837-6

Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219. https://doi.org/10.1145/3351095.3372860

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Colaner, N. (2022). Is explainable artificial intelligence intrinsically valuable? *AI & SOCIETY*, *37*(1), 231–238. https://doi.org/10.1007/s00146-021-01184-2

Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., & Bechara, A. F. (2023). Should explainability be a fifth ethical principle in AI ethics? *AI and Ethics*, *3*(1), 123–134. https://doi.org/10.1007/s43681-022-00152-w

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government information quarterly*, *39*(2), 1–8. https://doi.org/10.1016/j.giq.2021.101666

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 1–13. https://doi.org/10.48550/arXiv.1702.08608

Došilović, F. K., Brčič, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 210–215. https://doi.org/10.23919/MIPRO.2018.8400040

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, *24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines*, *28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1), 1–14. https://doi.org/10.1162/99608f92.8cd550d1

Franke, U. (2022). First- and second-level bias in automated decision-making. *Philosophy & Technology*, *35*(21), 1–20. https://doi.org/10.1007/s13347-022-00500-y

Freiesleben, T., & König, G. (2023). Dear XAI Community, We Need to Talk! Fundamental misconceptions in current XAI research. In L. Longo (Eds.), *Explainable Artificial Intelligence* (Vol. 1901, pp. 48–65). Springer. https://doi.org/10.1007/978-3-031-44064-9_3

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, *30*(3), 291–323. https://doi.org/10.1007/s10506-021-09294-4

Gyevnar, B., Ferguson, N., & Schafer, B. (2023). Bridging the transparency gap: What can explainable AI learn from the AI Act? In K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, & R. Rădulescu (Eds.), *ECAI 2023* (Vol. 372, pp. 964–971). IOS Press. https://doi.org/10.3233/FAIA230367

Hansson, S. O. (2017). *The Ethics of Technology: Methods and Approaches*. Rowman & Littlefield Publishers.

Hatherley, J., Sparrow, R., & Howard, M. (2026). The virtues of interpretable medical artificial intelligence. *Cambridge Quarterly of Healthcare Ethics, 35*(3), 167-176. https://doi.org/10.1017/S0963180122000305

Heidegger, M. (1977). *The question concerning technology and other essays*. Graland Publishing, Inc.

Herzog, C. (2022). On the ethical and epistemological utility of explicable AI in medicine. *Philosophy & Technology, 35*(2), 50. https://doi.org/10.1007/s13347-022-00546-y

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, *1*, 41–47. https://doi.org/10.1007/s43681-020-00008-1

Hildebrandt, M. (2020). *A philosophy of technology for computational law*. Oxford University Press. https://doi.org/10.31228/osf.io/7eykj

Hindocha, S., & Badea, C. (2022). Moral exemplars for the virtuous machine: The clinician's role in ethical artificial intelligence for healthcare. *AI and Ethics*, *2*(1), 167–175. https://doi.org/10.1007/s43681-021-00089-6

Ihde, D. (1979). *Technics and Praxis: A philosophy of technology*. D. Reidel Publishing Company.

Jongepier, F., & Keymolen, E. (2022). Explanation and agency: Exploring the normative-epistemic landscape of the "Right to explanation". *Ethics and Information Technology*, *24*(4), 1–11. https://doi.org/10.1007/s10676-022-09654-x

Kamath, U., & Liu, J. C. (2021). *Explainable artificial intelligence: An introduction to interpretable machine learning*. Springer International Publishing.

Karlan, B., & Allen, C. (2024). Engineered wisdom for learning machines. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*(2), 257–272. https://doi.org/10.1080/0952813X.2022.2092559

Kempt, H., Freyer, N., & Nagel, S. K. (2022). Justice and the normative standards of explainability in healthcare. *Philosophy & Technology*, *35*(4), 100. https://doi.org/10.1007/s13347-022-00598-0

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*, 487–502. https://doi.org/10.1007/s13347-019-00372-9

Lim, S. (2021). Judicial decision-making and explainable artificial intelligence. *33 Singapore Academy of Law Journal 280*, 280–314.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Mackenzie, D. (2006). *An engine, not a camera: How financial models shape markets*. MIT Press.

Maclure, J. (2021). AI, explainability and public reason: The argument from the limitations of the human mind. *Minds and Machines*, *31*(3), 421–438. https://doi.org/10.1007/s11023-021-09570-x

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, 1198–1212. https://doi.org/10.1145/3593013.3594074

Noorman, M., & Swierstra, T. (2023). Democratizing AI from a sociotechnical perspective. *Minds and Machines*, *33*, 563–586. https://doi.org/10.1007/s11023-023-09651-z

O'Hara, K. (2020). Explainable AI and the philosophy and practice of explanation. *Computer Law & Security Review*, *39*, 1–7. https://doi.org/10.1016/j.clsr.2020.105474

Peters, U. (2023). Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque. *AI and Ethics*, *3*(3), 963–974. https://doi.org/10.1007/s43681-022-00217-w

Ranchordas, S. (2021). Empathy in the digital administrative state. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3946487

Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Ed.). University of Chicago Press.

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Robbins, S. (2025). What machines shouldn't do. *AI & SOCIETY, 40,* 4093-4104. https://doi.org/10.1007/s00146-024-02169-7

Russo, F., Schliesser, E., & Wagemans, J. (2023). Connecting ethics and epistemology of AI. *AI & SOCIETY*, *39*, 1585–1603. https://doi.org/10.1007/s00146-022-01617-6

Scherer, M. (2019). Artificial intelligence and legal decision-making: The wide open? *Journal of International Arbitration*, *36*(5), 539–573. https://doi.org/10.54648/joia2019028

Searle, J. R. (2006). Social ontology: Some basic principles. *Anthropological Theory*, *6*(1), 51–71. https://doi.org/10.1177/1463499606061731

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Starke, G., Schmidt, B., De Clercq, E., & Elger, B. S. (2023). Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI and Ethics*, *3*(1), 303–314. https://doi.org/10.1007/s43681-022-00177-1

Taylor, I. (2024). Is explainable AI responsible AI? *AI & SOCIETY, 40,* 1695-1704. https://doi.org/10.1007/s00146-024-01939-7

Ursin, F., Lindner, F., Ropinski, T., Salloch, S., & Timmermann, C. (2023). Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? *Ethik in Der Medizin*, *35*(2), 173–199. https://doi.org/10.1007/s00481-023-00761-x

Vredenburgh, K. (2022). The Right to Explanation. *Journal of Political Philosophy*, *30*(2), 209–229. https://doi.org/10.1111/jopp.12262

Vredenburgh, K. (2023). AI and bureaucratic discretion. *Inquiry*, *68*(4), 1091-1120. https://doi.org/10.1080/0020174X.2023.2261468

Wasserman-Rozen, H., Gilad-Bachrach, R., & Elkin-Koren, N. (2023). Lost in Translation: The Limits of Explainability in AI. *SSRN Electronic Journal*, 1–42, https://doi.org/10.2139/ssrn.4531323

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy*, *5*, e2. https://doi.org/10.1017/dap.2022.39

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing* (pp. 563–574). Springer.

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

Ződi, Z. (2022). Algorithmic explainability and legal reasoning. *The Theory and Practice of Legislation*, *10*(1), 67–92. https://doi.org/10.1080/20508840.2022.2033945