Philosophy & Technology (2025) 38:4
https://doi.org/10.1007/s13347-024-00837-6

RESEARCH ARTICLE

Trust, Explainability and AI

Sam Baron[1]

Received: 31 July 2024 / Accepted: 24 December 2024 / Published online: 8 January 2025

Abstract
There has been a surge of interest in explainable artificial intelligence (XAI). It is commonly claimed that explainability is necessary for trust in AI, and that this is why we need it. In this paper, I argue that for some notions of trust it is plausible that explainability is indeed a necessary condition. But that these kinds of trust are not appropriate for AI. For notions of trust that are appropriate for AI, explainability is not a necessary condition. I thus conclude that explainability is not necessary for trust in AI that matters.

Keywords Explainability · Trust · AI · Reliability

1 Introduction

The use of artificial intelligence (AI) for decision-making has become widespread. This includes banking (Aggarwal, 2021), criminal recidivism (Dressel & Farid, 2018), hiring (Kuncel et al., 2014), policing (Alikhademi et al., 2022) and healthcare (Alowais et al., 2023; Gulshan et al., 2016; Donnelly, 2017; Longoni, Bonezzi and Morewedge 2019; McKinney et al., 2020; Topol, 2019). Many AI systems are opaque: no-one can explain why those systems provide the outputs that they do, either in specific cases or in general. Such opacity need not be not an all or nothing matter: while some systems are wholly opaque, others may only be opaque in certain ways or to some degree.[1] While my focus here is on wholly opaque systems, the discussion aims to be fully general, applying to any system that is opaque to some degree.

AI systems based on machine learning and deep learning techniques are often opaque, particularly those that implement a neural network. This is the type of AI system that I will focus on here, though, again, the discussion aims to be fully general: applying to any AI system that is opaque. My focus on such systems is primarily in

[1] See, for instance, Seddik et al.'s (2022) grey-box approach. For more on transparency, see Creel (2020) and Beisbart and Räz (2022).

✉ Sam Baron
s.baron@unimelb.edu.au

[1] Associate Professor of Philosophy, University of Melbourne, Melbourne, Australia

their use for decision-making, such as loan decisions, medical diagnosis or criminal recidivism prediction.

The lack of explainability for such systems is thought to pose a range of serious problems (Wachter et al., 2018). Of particular interest here is the relationship between explainability and trust. Trust in AI is thought to be important, as it is supposed that if AI systems are trustworthy then they are more likely to be used (Choung et al., 2021; Kelly et al., 2023; Ribeiro et al., 2016; Sullivan et al., 2022; Yang et al., 2023). Accordingly, if we don't build trustworthy AI systems then we may miss out on the substantial benefits offered from the widespread uptake of decision-making powered by AI.

Trust in AI is one of the primary motivations for explainability (Kästner et al., 2021). For if an AI system is not explainable, then it is thought it cannot be trusted. In other words, explainability is necessary for trust in AI. This idea can be found throughout the literature, as the quotes below demonstrate. The point is put in various ways: explainability is *needed* for trust; *in order to be* trustworthy AI must be explainable; explainability is a *pre-requisite* for trust; explainability is a *necessary condition* for trust and so on.

> In order for humans to trust black-box methods, we need explainability ... (Gilpin et al., 2018, p. 80)

> The need for explainable AI is motivated mainly by three reasons: the need for trust ... (Fox et al., 2017, p. 24)

> Artificial agents need to explain their decision to the user in order to gain trust ... (Pieters, 2011, p. 53)

> Explainability is ... a pre-requisite for practitioner trust. (Dam et al., 2018, p. 53)

> ... there is a need to explain ... so that users and decision makers can develop appropriate trust. (Mathews, 2019, p. 1271)

> In addition, model explainability is a prerequisite for building trust and adoption of AI systems (Gade et al., 2019, p. 3203)

> Therefore, society requires techniques in which XAI tools are an essential but insufficient step in determining whether or not an AI-based system can be trusted and employed for the task at hand. (Ali et al., 2023, p. 3)

> [trustworthy AI] should transparent [sic], that is, it should be explainable, traceable and communicable in a way that the creator and the user are able to understand the functioning of the AI. (Chamola et al, p. 78996)

> Defense is facing challenges that demand more intelligent, autonomous, and symbiotic systems. Explainable AI—especially explainable machine learning—will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners (DARPA 2016, quoted in Colaner, 2022, p. 233)

> A philosophical analysis of trust will show why transparency is a necessary condition for trust and eventually for judging AI to be trustworthy. (von Eschenbach, 2021, p. 1608)

But is explainability really necessary for trust? Answering this question is difficult, since there exist many accounts of trust. In this paper, I thus explore the relationship between explainability and these various accounts of trust. I argue that for some notions of trust it is plausible that explainability is indeed a necessary condition. But that these kinds of trust are not appropriate for AI. For notions of trust that are appropriate for AI, explainability is not a necessary condition. I therefore conclude that explainability is not after all necessary for *appropriate* trust in AI. For the kind of trust in AI that matters, there is no need for explainability.

The rest of the paper outlines this argument. In §2, I clarify the idea that explainability is necessary for trust by differentiating between different notions of trust. In §3, I consider notions of trust that are appropriate for AI and show that explainability is not needed. I then consider notions of trust for which explainability is plausibly necessary, and argue that they are not appropriate for AI. §4 considers some objections to the main argument. I finish, in §5, with some general reflections on trust and explainability.

Note that parts of this argument have been made already by Ferrario and Loi (2022), Ryan (2020) and Frieman (2023). However, these works tend to focus either on the appropriateness of trust, or on the relationship between explainability and trust for only certain types of trust. For instance, Ferrario and Loi (2022) argue that certain notions of trust based on reliance do not require explainability. However, they do not consider the broader space of approaches to trust and the connections they bear to explainability. Similarly, Ryan (2020) and Frieman (2023) argue that some notions of trust are inappropriate for AI, but do not consider how this relates to explainability.

The present paper thus goes beyond this existing work in four ways. First, the paper considers the relationship between explainability and trust for notions of trust where this relationship has not yet been considered. Second, the paper looks at notions of trust in AI that have not previously been considered (e.g., Holton's, 1994; Nickel's, 2013, 2022; and Nguyen's, 2022 approaches). Third, the paper takes a deeper look at whether reliance requires explainability by considering ways in which one might see explainability as necessary for the reliability of AI systems. Finally, the paper draws these strands on trust and explainability together, by providing the first general argument against the idea that explainability is needed for trust in AI that matters.

2 Types of Trust

In this section, I will introduce several accounts of trust. Trust in each case is to be differentiated from mere reliance. For instance, one relies on a hammer not to break, or a ladder to carry one's weight. In this case, one is relying on a specific tool to perform its function. Reliance is usually treated as necessary but not sufficient for trust. Trust is reliance plus something extra (Baier, 1986; Jones, 1996; Holton, 1994; Goldberg,

2020; Nickel, 2013, 2022; Nguyen, 2022; McLeod, 2002; von Eschenbach, 2021). There are seven distinct accounts of this further feature available.

The first is Nguyen's (2022) conception of trust as an unquestioning attitude. This account of trust can be stated as follows:

> To trust X to P is to have an attitude of not questioning that X will P (Nguyen, 2022, p. 225).[2]

The unquestioning attitude is further explained in terms of two dispositions. First, the disposition to immediately accept that X will P. Second, a higher-order disposition to "deflect questioning" about the first disposition. So, for instance, Ping trusts Simon to look after her cat when Ping immediately accepts that Simon will look after her cat and, moreover, Ping does not question this first-order disposition to accept. Thus, Ping does not ask herself should she *really* let Simon look after her cat? Should she ensure that José looks in on Simon? Should she send extra biscuits with her cat just in case Simon forgets to buy food? And so on. The idea, in essence, is that Ping trusts Simon when she accepts that he will do this thing and she does also does not give it a second thought.

Importantly, the unquestioning attitude can be adopted toward both people and things. For instance, Ping can trust her computer in the relevant sense, when she accepts that her computer will save her document, and doesn't spend any time questioning that acceptance. She doesn't for instance, repeatedly hit save, or google how save works to make sure her computer will really do the job, nor does she seek to better understand the capacities of her computer and so on. She can of course do all of these things. The point is that when she trusts her computer, she doesn't do any of them.

A second, closely related, notion of trust is advocated by Ferrario and Loi (2022). This account of trust is explicitly offered as an account of trust for AI. The idea, very roughly, is that trust in AI is reliance without needing to continue checking that the AI is doing its job well. This account shares some obvious similarities with Nguyen's account, but there appear to be some subtle differences. For one thing, Nguyen's account appears to be a bit broader. The disposition to accept without question is, presumably, held against all types of questioning, not just questions associated with checking to make sure some object is performing its function well. For another thing, Ferrario and Loi's account is not offered as a general account of trust, unlike Nguyen's.

The third account of trust is Nickel's (2013) entitlement account. According to this account, trust is a matter of relying on the performance P of someone or something where (i) it is worth relying on P and (ii) we are entitled to rely on P. Note that the 'entitlement' is a normative expectation, specifically a normative expectation about performance P. For instance, suppose once again that Ping trusts her computer to save her document, where this is sufficiently valuable that it is worth relying on her computer to do that. This satisfies condition (i). Her reliance counts as trust when she is entitled to rely on her computer to save the document, in this sense: she expects the computer to save the document (the expectation part) and it *should* (the normative

[2] Strictly speaking, this statement of the account (which is Nguyen's) does not mention reliance. However, he makes it clear that "to trust is to rely on a resource while suspending deliberation over its reliability" (Nguyen, 2022)

part). The 'should' here can be understood as something like proper functioning: part of what the computer is *for* is processing and saving documents. If that is what the computer is for, then (so the thought goes), it *should* perform that task.

In a recent paper, Nickel (2022) offers an account of trust for AI, involving *discretionary authority*. The idea is that we trust AI in the sense that we give it discretion over something we value, where that discretion is based on normative and predictive expectations about the AI system in question. To give something discretion in a specific domain appears, very roughly, to involve deferring to that thing for something we value. So, for instance, a clinician might defer to an AI system for diagnosis, which is something of great value to the clinician. In so doing, however, the clinician may have expectations of the AI. They might expect that the AI produces a diagnosis (a predictive expectation) and that the AI does its job well (a normative expectation), insofar as the diagnosis is accurate.

The discretionary authority account appears to be broadly an application of the entitlement account to the case of AI. That being said, there are some differences in the two pictures. For the discretionary authority account, Nickel discusses a role for the AI practitioner—those who design and generate the AI. Moreover, the AI practitioner can be morally accountable in cases where discretionary authority is given to an AI and the AI system fails. This added moral dimension is used to explain the moral inflection to trust in AI without imbuing the AI itself with moral obligations. For now, I will elide the difference between the discretionary authority account and the entitlement account, as the difference does not matter for my argument.

The fourth view is an affective notion of trust. The broad idea behind this account is that trust is a matter of believing that the trustee holds some positive affect toward the trustor. On a simple version of this account, x trusts y when x relies on y and x believes that y will act of out of good will toward x (Baier, 1986; Jones, 1996). So, for instance, Cathy trusts Ping to look after her cat Simon, when she relies on Ping to do so and believes that Ping will look after Simon out of goodwill toward Cathy. If Cathy doesn't believe that Ping will look after Simon out of goodwill toward Cathy, she won't trust Ping in the relevant sense.

The fifth, related approach focuses on normative, rather than affective motivations (McLeod, 2002). On a simple account of normative trust, x trusts y when x relies on y and x believes that y will act out of a normative commitment toward x. Thus, Cathy trusts Ping to look after Simon when she relies on Ping to do so and believes that Ping will look after Simon because she is motivated to do the right thing, morally speaking, by Cathy.

The sixth view is the 'trust-responsiveness' account (Jones, 2012; Hawley, 2014). On this approach, the trustor holds an expectation regarding the trustee to act with a responsiveness to the right reasons. Thus, Cathy trusts Ping to look after Simon, when Cathy relies on Ping to do so and, moreover, believes that the fact Cathy relies on Ping weighs on Ping in their decision-making. Thus, Ping takes into account, as a reason for choosing a particular course of action, Cathy's reliance on them. Note that this is compatible with Ping having a wide variety of motivations to act, or indeed with no particular motivation. Ping may be motivated to act out of goodwill or a normative commitment, but she may also not be so motivated.

Of course, presumably Ping must have some motivation to act or other. The point, though, is that Cathy's trust is not based on believing Ping has any specific motivation,

but rather is based on believing that Ping is responsive to reasons of the right kind in decision-making. To highlight the point, Ping could have only ill-will toward Cathy, and still Cathy could trust Ping because she knows Ping will take Cathy's reliance seriously, even though Ping hates Cathy (*mutatis mutandis* for pretty much any other motivation one might attribute to Ping).

The seventh and final notion of trust is Holton's (1994) decision-based account. Roughly speaking, the account has two parts. First, for Holton, trust is based on a decision to trust. Importantly, one can decide to trust even in cases where one knows that whomever one is trusting has no good will, or normative commitment. One could, for instance, decide to trust a new employee because doing so is part of giving them a chance, perhaps because their options are limited. One might do this despite not believing that the employee acts out of good will, or out of a normative commitment. Rather, one sees an additional source of value in extending trust in this case, and one's decision to trust aims to secure that value.

Second, when one decides to trust some x, this is partly a matter of adopting a particular stance toward x: the participant stance. Adopting the participant stance is a matter of having a readiness to adopt certain reactive attitudes, such as betrayal and gratitude. Together, with reliance, adopting the participant stance is constitutive of trust on Holton's view. As he puts it:

> When you trust someone to do something, you rely on them to do it, and you regard that reliance in a certain way: you have a readiness to feel betrayal should it be disappointed, and gratitude should it be upheld. (Holton, 1994, p. 67)

So, for instance, suppose that Cathy trusts Ping to look after her cat Simon. When Cathy does this, she is deciding to rely on Ping, and she readies herself to have certain reactive attitudes. In particular, if Ping fails to look after Simon then Cathy is likely to feel betrayed, and if Ping succeeds then Cathy is likely to feel grateful. The motives behind Ping's actions, on this view, are not obviously relevant. If Ping succeeds in looking after Simon but without any goodwill or any moral commitment to Cathy, then Cathy is still likely to feel gratitude in virtue of having adopted the participant stance toward Ping.

2.1 Back to Explainability

These seven approaches to trust give us seven ways of clarifying the idea that explainability is necessary for trust. How should we understand explainability with respect to each of these cases of necessitation? For present purposes, I will focus on local explainability. Explainability in this sense is a matter of providing understanding of why an AI system produces a particular output, on a particular occasion. So, for instance, suppose that Ping applies for a loan, and all of their information is fed into a machine learning model that determines their credit score. The model delivers a low credit score for Ping. In response, Ping can demand an explanation: they can ask why it is that the algorithm delivered a low credit score on this occasion. Local explainability is a matter of giving Ping an answer to this question.

I focus on local explainability simply because it has received a great deal of attention, with a number of approaches to local explainability developed already. Chief among these is the counterfactual approach, according to which giving an explanation is a matter of providing a counterfactual linking the output of an algorithm to its inputs (Wachter et al., 2018). So, for instance, on this approach Ping's low credit score is explained by their income when, had their income been higher, they would have achieved a higher credit score. Note that while I focus on local explainability, I return to global explainability in §4.

A further point of clarification: it is important to pull apart the *possibility* of trust in AI from the *appropriateness* of trust in AI. Trust, in all of the senses just described, is a matter of the trustor adopting a particular mental state toward the trustee: an attitude, a belief, a stance—whatever. Trust in AI is therefore possible insofar as it is possible for someone to enter into the relevant mental states with AI as the subject. While it is tempting to think that trust in AI in some sense is impossible—because AI doesn't have, say, normative commitments, a point I return to below—this would be a mistake. There is, in general, no reason at all to suppose that people can't have mental states toward AI that qualify as states of trust. It is, in general, open for any of us to adopt pretty much any belief or attitude about anything, and AI is no exception.

The appropriateness of trust in AI is a different story. Roughly speaking, the appropriateness of trust in AI corresponds to the fittingness of the relevant trusting mental state one has toward AI. So, for instance, Nguyen-style (2022) trust toward AI is appropriate if it is in fact fitting to adopt an unquestioning attitude toward AI. For a clear example of when it is not fitting to trust, consider a case in which one trusts a hammer in the normative sense of trust (McLeod, 2002). Relying on a hammer and believing that it acts out of a normative commitment would clearly be a mistake, twice over: hammers don't act in the relevant sense, and they lack anything like the advanced capacities needed for either motivation or motivation out of commitment to a moral principle.

The question we must ask, then, is not whether one *can* trust AI in any of the senses outlined above, because clearly one can (on pain of bizarre constraints on which mental states one can have). Rather, the question we must ask is whether it is appropriate to do so and, if so, whether explainability is necessary for appropriate trust in the relevant sense.

3 Trust and Explainability

We are interested in whether explainability is necessary for appropriate trust in AI. In what follows, I take each notion of trust and consider this question by looking at (i) whether trust in the relevant sense is appropriate and (ii) whether explainability is necessary for the kind of trust at issue. First, however, I will consider the notion of reliability. I do this because, as noted, reliance is necessary for each of the seven notions of trust canvassed above. I thus focus on appropriate reliance, since it is appropriate trust that is at issue. If appropriate reliance necessitates explainability, then this would show that all seven notions of trust necessitate explainability, thereby undermining my argument. I will thus argue that appropriate reliance on AI does

not require explainability. This will pave the way for the subsequent discussion of explainability and trust.

3.1 Reliance

We can, and clearly do, rely on AI. Moreover, in many cases this reliance appears appropriate. Relying on something for a particular purposes is appropriate when we have good evidence that whatever we are relying on works well for the intended purpose. What it is for something to work well for an intended purpose really depends on the purpose at issue.

Take a hammer, for example. A hammer is reliable when it performs its function well, over time. In general, we can say that the hammer is reliable when it meets a set of expectations that we have for it, based on the task at hand. These expectations include fulfilling its purpose (driving nails) but also not breaking easily, being comfortable to use for long periods of time and so on. Whatever it is, in short, that allows for the task at hand, and similar tasks over time, to be completed well, and without the hammer itself becoming particularly salient in the context of such tasks by, for instance, breaking down or otherwise being a nuisance.

For AI, reliance seems appropriate when a given AI system produces highly accurate outputs over time.[3] For example, reliance on AI for medical diagnosis is appropriate, when it regularly offers accurate diagnoses. Similarly, reliance on AI is appropriate when it regularly produces highly accurate credit ratings. Clearly there will be differences between the diagnosis and credit scoring cases. Whether and to what extent AI is reliable depends on the task, the system and a range of other factors.[4]

Still it is plausible that at least some AI tools can be relied upon because of their high degree of accuracy over a number of uses, and thus that reliance is at least sometimes appropriate. This hardly seems controversial: AI is a tool and sometimes it is appropriate to rely on that tool. The question we need to answer is whether explainability is necessary for appropriate reliance on AI in the relevant sense. The answer seems to be 'no'. That's because explainability does not seem to be necessary for mere reliance in general. To show this, I will begin with some general reflections on reliance and explainability for cases that don't involve AI before considering the AI case.

We rely on a great many things without having the faintest idea of how they work, or why they do what they do. So, for instance, suppose that Ping buys a new computer. Ping comes to rely on their new computer, and it becomes a core tool in their job. But Ping has no idea why the computer does what it does, or how it does what it does.

[3] On statistical accuracy in this context, see Belkin et al. (2019) and Grote et al. (2024).

[4] We can sharpen this as *computational reliabilism* (Durán & Formanek, 2018; Durán & Jongsma, 2021). Roughly: it is fitting for x to rely on an AI when it regularly produces accurate results through a reliable process (e.g., the algorithm). Durán and Formanek (2018) give four indicators of process reliability. The details need not concern us. What matters is that reliability under computational reliablism does not presuppose explainability (Durán and Jongsma, 2021, p. 332). Computational reliabilism is controversial and so, below, I consider another approach involving modal robustness.

Ping can rely on their laptop because they can use it successfully without possessing any real explanatory information.[5]

What happens when Ping's computer breaks down? Surely, this points to the need for some understanding of how the computer works or why it does what it does. But this doesn't seem all that plausible. Suppose Ping knows that the laptop could break down, but that it does so very rarely and, when it does break down, they can get it fixed. In this situation, they seem to have enough information to rely on the laptop. Their expectations of the laptop, given the task that they are trying to perform, are met.

Still, one might worry that some notion of explainability is needed. For Ping to know that their laptop won't break down easily, and for their laptop to be fixed when it does breakdown, someone needs to know how the laptop works. Thus for Ping to truly rely on the laptop, it must be explainable in some sense to someone, even if not to Ping.

Note, however, that even if no-one understands how Ping's laptop works well enough to fix it when it breaks, Ping could still come to rely on it. Simply using the laptop successfully for a long period of time seems enough for Ping to reasonably rely on the laptop. This is because, through repeated use, Ping can gain enough information about the laptop's past behaviour to run a reasonable induction on how it will behave in the future; reasonable enough for them to rely on the laptop for the task at hand.

Of course, explainability does tend to promote reliance. If Ping knows that someone knows how their laptop works, and how to fix it if it is broken, then Ping can be more confident in their laptop. They can rely to a greater extent on the laptop, for a greater range of tasks, or perhaps for a longer period of time.

That explainability tends to promote reliance becomes even clearer once one recognises that understanding how something works puts one in a position to make it better, or more fit for purpose. Thus, suppose that Ping has a laptop that does not work very well. They rely on it, we may suppose, but not very much. If they know how the laptop works, they can perhaps troubleshoot it in order to make it work better and perhaps more stably over time.

Still, it remains the case that explainability is not necessary for reliance. Knowing that an object can in fact meet one's expectations for a particular task is enough to rely on it. What's key is that one can know that something meets one's expectations for a particular task without knowing any explanatory information. That is because one can carry out 'experiments' involving the object in question, to see if it works well enough for a given task. Do enough of these experiments, and one will build up a sufficiently solid inductive base concerning the capacity of an object to function for a specific purpose to form a basis for reliance.

At this point, one might be tempted to draw a distinction between explainability in principle and explainability in practice. It seems plausible that Ping's laptop is explainable in principle, even if in practice Ping can't explain it. One might argue that while being explainable in practice is not necessary for reliance, being explainable in

[5] Alonso (2016) outlines several accounts of when reliance is appropriate. On none of the accounts is explanatory information necessary for reliance in any obvious sense.

principle is. If no-one can explain how Ping's laptop works, then Ping has no business relying on it.

But this, again, is implausible: so long as Ping can have a reasonable belief that the laptop meets their expectations, including functioning for a particular purpose stably over time, it doesn't matter whether it is explainable even in principle. Suppose the laptop is some alien device. If, after using it for a long time, Ping finds that it doesn't break down and can perform the tasks required of it, then it seems appropriate for Ping to come to rely on the device. Mere reliance on something just requires evidence of a certain kind of functioning. While explainability can be a part of the relevant evidential base, it is not a necessary part of that base. Again, this is all compatible with explainability promoting reliance on some tool.

We can now carry these general considerations over to the case of AI. Suppose that we have good evidence that an AI system produces highly accurate predictions with regard to credit scoring over a large number of instances. This evidence alone seems to be sufficient for relying on the AI for credit scoring. Do we need explainability to gather this evidence? It doesn't seem so. The methods for gathering this evidence and thus establishing the reliability of an AI system don't seem to require explanatory information. That's because these methods tend to be statistical or inductive in nature and such methods don't typically presuppose an understanding of how a system works (see e.g., Belkin et al. (2019); Durán and Formanek, 2018; and Grote et al., 2024 for discussion).[6]

As Duede (2023, pp. 1097–1098) puts the point "inductive considerations are... sufficient to establish reliability" in the case of AI systems and the outputs of such systems "can be independently verified, thereby rendering opacity epistemically irrelevant" (Duede, 2022, p. 491, see also Ryan, 2020 p. 11). Opacity thus does not prevent reliability (Durán and Jongsma, 2021, p. 332).

As in the more general case of reliance, explainability can promote reliance on AI (see, e.g., Scharowski et al., 2023; Schemmer et al., 2023; Schoeffer et al., 2024). One obvious way in which this can happen is any case in which making a system explainable helps to improve its accuracy. In this situation, we can use explainability directly to lift an under-performing AI over some accuracy threshold, so that it can then be relied upon for a particular task. But, as before, explainability is not necessary for increasing the predictive accuracy of AI. There are other ways to do this: for instance, one can also improve data quality (Hong et al., 2023).

What about explainability in principle versus explainability in practice? Does that make a difference for appropriate reliance on AI? One might think so. Suppose that a credit scoring system is not explainable in principle. Then, one might argue, it is not appropriate to rely on that system. But, again, this seems implausible. So long as we know that the system is predictively accurate, it does not seem to matter whether the system is even explainable in principle. Reliance on the system requires only that we can be confident that it works, and this confidence does not seem to require of us that we know how it works.

[6] That said, it remains somewhat controversial as to what reliability involves for an AI system. For instance, as Grote et al. (2024) note, Duede (2022) seems to require explainability for some cases. See Sullivan (2023) for an approach to reliability that does not rely on explainability that candle Duede's cases.

I have argued that mere reliance on AI does not require explainability either in principle or in practice. That's because it is appropriate to rely on AI so long as we have strong evidence that it is predictively accurate. One might object, however, that predictive accuracy is not sufficient for reliance on AI to be appropriate. In addition to predictive accuracy, we need information about modal robustness.[7] That is, we need to know that the AI system will continue to be predictively accurate under various changes to the context in which it is put to use. But, one might argue, the only way to gain information of that kind is through explainability. We need to know how an AI system produces a particular output to know whether it would continue to be accurate if circumstances change.

But, again, it is far from clear that this is true, for two reasons. First, it is not clear that modal robustness of the relevant kind is strictly necessary for appropriate reliance on AI. If, for instance, one has good evidence that the circumstances won't change, or one has no intention of using the AI in a situation where circumstances are subject to change, then predictive accuracy seems to be enough. Second, it is possible to gain information about modal robustness without explainability. As before, one can do this by experimenting with the AI in different circumstances, measuring its outputs for accuracy in those circumstances (Freiesleben & Grote, 2023; Grote et al., 2024). None of this requires knowing why the AI system produced a given output.

In sum, then, mere reliance does not seem to necessitate explainability. Explainability can, of course, promote reliance, and so there is still perhaps a role for explainability to play in supporting appropriate reliance, but it is by no means necessary for reliance, and it seems we can get along perfectly well without it so long as we have good information about the accuracy and perhaps modal robustness of an AI system.

It is important to be clear about the scope of this conclusion. The claim I have been arguing for is this:

> **T1:** Explainability is not necessary for appropriate reliance on AI, in cases where good information indicating the reliability of an AI system is available.

There is, however, a weaker claim available, namely:

> **T1*:** Explainability is important for fostering appropriate reliance on AI in cases where good information indicating the reliability of an AI system is not available.

Everything I have said leaves T1* open. Thus, it may be that in cases where we do not have good information indicating the reliability of an AI system, explainability is an important strategy for gathering that information. Indeed, given that it can be quite difficult to establish the reliability or modal robustness of an AI system, it is reasonable to suppose that explainability has a role to play. It seems doubtful, however, that explainability is *necessary* for gathering the relevant information in general. That's because, as discussed, establishing the reliability of an AI system does not seem to require explainability.

[7] Freiesleben and Grote (2023) analyse reliability in terms of robustness. As Grote et al. (2024) discuss, the relevant notion of robustness is increasingly linked to modality. See, for instance, Buijsman (2023), Vandenburgh (2023).

3.2 Moderate Trust

We come now to the seven notions of trust outlined in §2. To break up the discussion a bit, I will group these seven approaches to trust into 'moderate' and 'strong' accounts. As noted, for all notions of trust considered, trust is reliance plus something extra. Strong notions of trust generally take the added feature to involve another agent. The broad idea being that trust is an interpersonal notion, with various ways of filling out the interpersonal dimension. Moderate notions of trust, by contrast, take the added feature to not involve another agent. Strong notions are thus strong because they are more demanding. They require other agents in the mix; not so for moderate notions of trust.

I take Nguyen's unquestioning attitude account, Ferrario and Loi's monitoring account and Nickel's entitlement account to be moderate accounts of trust. I will consider these as a group, before turning to strong notions of trust.

As we saw, Nguyen's notion of trust involves the adoption of an unquestioning attitude toward some x. By adopting this attitude one relies on x without subjecting x to further critical examination. For instance, when a climber relies on a rope without thinking about whether they should do so, or whether the rope might break, they are adopting an unquestioning attitude. Nguyen eventually frames this in terms of the extension of agency. The idea is that one takes whatever one relies on to be part of one's means of exerting agency in the world. As Nguyen puts it "my car, my mouse, my laptop—these objects have come to be functionally integrated with me to various extents" (Nguyen, 2022, p. 231). When something integrated in this way breaks down, this can signal a breach of trust.

Is it ever appropriate to trust AI in Nguyen's sense? The answer appears to be 'yes'. In many cases, the adoption of an unquestioning attitude is appropriate when one has 'tested' whatever it is one aims to trust, and that thing has passed the test. So, for instance, a climber might repeatedly test a rope, until they are satisfied with its capabilities. Having done so, it is fitting to use it and to thus adopt an unquestioning attitude. Tests of this kind can be easily conducted for an AI system as well. Granted, not every AI system will pass the test, and so an unquestioning attitude may not always be appropriate. But there is nothing that fully blocks the appropriateness of this attitude for AI, at least not that I can see.

The question, then, is whether explainability is necessary for appropriate trust in AI in Nyugen's sense. Well, notice that explainability is not necessary for the appropriate adoption of an unquestioning attitude in general. We appropriately adopt an unquestioning attitude to many things without making any demands for explainability. For instance, I adopt this attitude toward a car, even though I have at best a faint (and probably largely wrong) understanding of how it works. Similarly, I don't really understand how my laptop works and yet I am unquestioningly relying on it right now.

These cases of trust seem entirely appropriate, despite the absence of explanatory information. This is largely because one can appropriately adopt an unquestioning attitude based just on an induction over the past success of an object to function as desired, where this induction leads one to extend one's agency over the object. For instance, in the case of a climber and their rope one can "trust this rope to hold [one's]

weight because it has held it so many times in the past" (Nguyen, 2022, p. 227) at which point the rope can be integrated into the climber's agency. Importantly, one need not understand why the rope holds one's weight, or how ropes in general hold weight, or anything about the physics of ropes. One just needs a sound inductive basis to draw upon.

Is there then something special about the case of AI that makes explanatory information necessary? It seems doubtful. The appropriate extension of agency via the adoption of an unquestioning attitude toward AI does not seem to require an understanding of how the AI works, just an understanding that it does work. This understanding can be gained by drawing on the same inductive base that we use to establish the reliability of the AI system.

Similar considerations apply to Ferrario and Loi's (2022) monitoring account. As discussed, on this picture trust in x is a matter of relying on x without monitoring x. Trust as reliance plus monitoring is clearly appropriate for AI. One can reasonably rely on an AI system, without needing to constantly monitor it. Moreover, this is likely to be an important kind of trust for many AI systems, since in many cases we may wish to run these systems without monitoring them, particularly when constant monitoring is expensive or otherwise impractical.

The question, then, is whether explainability is necessary for appropriate trust in AI in this sense. Again, the answer is 'no'. For trust in AI of this type to be appropriate, one must be in a position to know that the system works well without constant supervision. It seems entirely possible, however, to gain information of this type without understanding why the system produces any given output. Again, the notion of experimentation with a system comes into play. We can simply test the AI on a sequence of cases and see how well it does, both for a given output and over time. If the system works stably and accurately, then we can rely on it without supervision.

It is also not clear how explainability would help with monitoring. Knowing why a system produces a particular output does not seem to provide much help with regard to understanding the stability of the system over time. Indeed, this is precisely what the proponents of the monitoring approach have recently argued: explainability is neither necessary nor sufficient for monitoring trust (Ferrario & Loi, 2022).

This brings us to Nickel's entitlement account. On this picture, trust in AI amounts to having a normative expectation that an AI system will do what it should. This 'should' does not signal moral normativity in the first instance but, instead, relates to the functioning of the system. An AI system might be designed to perform a specific function. In virtue of being so designed, one can trust the system in this sense: one can rely on the system for that function, and reasonably expect the system to perform the specified function. One's trust is fitting, presumably, when those normative expectations are appropriate and satisfied: when the system is in fact supposed to perform a certain function and does in fact do what one expects it to do. Trust in AI in this sense is breached if the AI fails to perform the function that is expected of it.

As with the two other notions of moderate trust, Nickel's account appears appropriate for AI. One can appropriately trust AI in this sense, because one can reasonably have the expectations that Nickel identifies, and those expectations are capable of being satisfied. AI systems are generally designed for a specific task, such as medical diagnosis, or credit scoring, and this generates normative expectations surrounding

the way those systems function. Those expectations can also be satisfied by virtue of the discovery that the AI system does in fact function as intended, by providing (say) highly accurate diagnoses of medical conditions.

As with other forms of moderate trust, explainability does not seem to be required for Nickel-style trust to be appropriate. As noted, for one's trust in this sense to be appropriate, one's expectations need to be satisfied. One can come to know that they are, however, without knowing how an AI system works, or why it produces the outputs it does. So long as one knows that the system reliably does what it is supposed to do, then that's all one needs for appropriate trust in the relevant sense. This information can, however, be gathered in just the same way that we determine the system's reliability. By using it, and observing how it operates. Over time, we will be able to determine that the system is doing what it should do, even if we don't know why.

3.3 Strong Trust

I turn now to the three stronger notions of trust outlined in §2. The first two notions of trust to consider are the affective account (Baier, 1986; Jones, 1996) and the normative commitment account (McLeod, 2002). These two accounts can be considered together. For both accounts, it is quite plausible that explainability is a necessary condition on trust in the relevant sense.

To see this, consider a non-AI case. As before, Cathy trusts Ping to look after her cat Simon. For normative and affective notions of trust, Cathy needs to understand why Ping acts in order for the trust to be appropriate. Take, as an example, the affective notion of trust. In order for Cathy's trust in Ping to be appropriate, Cathy needs to know that Ping acts because they are motivated by goodwill. In this way, when Ping looks after Simon, Cathy needs an explanatory understanding of Ping's actions. Without this understanding, it is difficult to see how her trust in Ping can be appropriate. Similar considerations apply to a normative notion of trust. In order for Cathy's trust to be appropriate, she needs to understand why Ping acts. In particular, she needs to know whether Ping acts out of a moral commitment to her, and so some explanation of Ping's actions is required.

Like the normative and affective notions of trust, explainability seems required for the trust-responsiveness account defended by Jones (2012) and Hawley (2014). In order for Cathy to trust Ping, she needs to know that Ping's decision-making is responsive to reasons, in particular that Ping's decision to look after Simon is based, in part, on the fact that Ping feels the weight of Cathy's reliance upon them. In this case, Cathy needs at least some explanatory information. Specifically, she needs to know why Ping makes the decision they do, and thus whether or not Ping is sensitive to the right reasons.

For Holton's (1994) notion of trust, the importance of explainability seems less obvious. Recall that, on Holton's view, trusting x is a matter of relying on x plus adopting a participant stance toward x. We've already seen that explainability is not necessary for reliance. The question, then, is whether it is necessary for adopting the participant stance. It doesn't seem to be. On the face of it, one can choose to adopt the participant stance to some x without understanding anything about why x behaves the

way it does. So, for instance, Ping could choose to adopt the participant stance toward Cathy, thereby adopting a readiness to have certain reactive attitudes like betrayal or gratitude, without really understanding much about why Cathy does what she does.

That said, it really depends on what is involved in adopting the 'participant stance'. For it could be that adopting this stance requires knowing that the trustee makes decisions in a certain way, in the manner that a person would. In this case, some explanatory information of the type needed by the other strong accounts of trust may well be required. At times, Holton's discussion tends in this direction. He seems to think the reactive attitudes are simply not held toward machines. As he puts it, "...when a machine breaks down we might feel angry or annoyed; but not (unless we are inveterately anthropomorphic) resentful" (Holton, 1994, p. 66). The parenthetical is interesting here. What it suggests, is that the reactive attitudes associated with trust, like betrayal, won't arise unless one anthropomorphises a machine, suggesting that beliefs about how the AI operates—perhaps that it makes decisions in a human-like way—would be needed to adopt the participant stance toward it. This could speak in favour of explainability as a condition on Holton-style trust.

Tentatively, then, we may say that explainability is necessary for at least some strong notions of trust. The question, then, is whether trust of this kind is ever appropriate for AI. The answer seems to be 'no'. This is relatively straightforward for the affective, normative and trust-responsiveness notions of trust. Each of these stronger notions of trust is appropriate when one can reasonably believe that the trustee has certain mental states or properties. These include: goodwill, moral commitment and responsiveness to reasons.

So far as we know, AI systems of the kind currently used to predict credit scores, or to undertake medical diagnosis do not have mental properties (Butlin et al., 2022). As Ryan (2020) notes "artificial agents do not have emotions or psychological attitudes for their motives, but instead act on the criteria inputted within their design or the rules outlined during their development" (see also Nahmias et al., 2020; Taddeo 2010, 2011). Even the large language models used in generative AI—that have been the subject of some scrutiny on this issue—lack these properties (Chalmers, 2023). As von Eschenbach (2021, p. 1613) puts it with respect to affective notions of trust, trust "would not be fitting due to the absence of evidence that these systems would carry out these functions in a manner consistent with our goods or interests". This extends to normative and trust-responsiveness accounts as well: AI systems do not have moral commitments and are not responsive to reasons because they lack mental properties of the right kind.

What of Holton's (1994) approach? For trust of this kind to be appropriate, the reactive attitudes of betrayal or gratitude would need to be fitting. But these attitudes do not seem to be fitting for AI. Take betrayal. It doesn't seem appropriate to feel betrayed by AI because it doesn't seem to be the kind of thing that can betray.[8] For betrayal to be a fitting response to x, x seems to require some degree of mentality.

[8] Nguyen (2022) does talk about being betrayed by an object, when that object is integrated into one's sense of agency. But, as already discussed, it is doubtful that Nguyen's approach requires explainability and, at any rate, it is plausible that Holton and Nguyen are interested in different, but perhaps related, notions of betrayal.

Exactly what kind of mental state is required for betrayal to be fitting is unclear. However, one possibility is suggested by Stout:

> X betrays Y with respect to a relationship between them that is partly constituted by the commitment to be loyal when the trust that Y has in including X in the relationship is breached by X failing to maintain this loyalty. (Stout, 2022, p. 340).

Rachman offers a broader account, stating that:

> Betrayal is a sense of being harmed by the intentional actions, or omissions, of a person who was assumed to be a trusted and loyal friend, relative, partner, colleague or companion. (Rachman, 2010, p. 304)

On either account of betrayal, feeling appropriately betrayed by a hammer doesn't obviously make sense because hammers are not psychologically sophisticated enough to betray. Of course, one might disagree with these accounts. But even if they are not exactly right, the broad idea seems plausible: *x*'s attitude that *y* has betrayed them seems fitting only if *y* has consciously chosen a certain course of action in some sense, thus making the possession of mental features necessary for betrayal. That being so, AI cannot be the betrayer in Holton's sense, since it lacks the right mental features.

In sum, it is hard to see how any stronger notion of trust could be fitting for AI, since AI systems lack the right mental properties. Perhaps this will change. If, in the end, we develop generalised artificial intelligence, which essentially has human level cognitive capacities, then perhaps it will be appropriate to have trust in AI in some strong sense. But that is simply not where we are at right now, and so trust in AI in any strong sense remains inappropriate.

4 Objections

This concludes my argument. To summarise the argument: while moderate notions of trust are appropriate for AI, it is plausible that explainability is not a necessary condition on those notions. For strong notions of trust, by contrast, explainability likely is a necessary condition, but such notions are not appropriate for AI. I anticipate three objections.

4.1 Objection One

First objection: I have focused on trust in AI systems. But, one might argue, taking a broader perspective will yield a different result. Instead of focusing on trust in AI systems in isolation, we should focus on AI systems in combination with those who employ them (Ferrario & Loi, 2022). So, for instance, in a medical context, rather than focusing on trust in AI, we should be focusing on the combination of a physician using an AI system and the system itself. We might even take a broader perspective still, and include a landscape of individuals involved in the development of AI as well as individuals like doctors who use those tools.

We can of course take this broader perspective. Once we do, stronger notions of trust come into play. Take the case of a doctor using an AI tool for diagnosis. Because the doctor is capable of having mental states, the conditions for, say affective rust are present. Thus, one can appropriately trust a doctor to use an AI system for diagnosis, where this involves an appropriate belief that the doctor uses the tool with goodwill toward the patient. Similarly, one can take the 'participant stance' toward the doctor, or one can reasonably believe that the doctor acts under the weight of the patient's reliance upon them.

Not only is it the case that one can appropriately trust in this stronger sense once one considers the doctor using AI, it also seems likely that trust in this sense is important. We really should be able to trust doctors to use AI tools with the right motivations, and with the patient's interests weighing on them in their deliberations. Moreover, given what I have argued above, explainability appears necessary for trust in this stronger sense.

If we take this broader perspective, then explainability does indeed seem important, but not explainability of the AI system. Rather, what one needs to understand is why the doctor acts: do they act from goodwill? From a normative commitment? In a manner that is responsive to the patient's reliance on them? These questions are about the doctor's mental states, and have little to do with the specific tools that they are using. That being so, adopting this broader perspective still does not bring with it a need for explainability of the AI system.

One might demur: it is necessary for the doctor to act out of goodwill toward the patient (say) that they understand why the tools they are using work. Thus, in order to reasonably believe that the doctor has goodwill, one must also believe that the doctor is in possession of an explanation for the particular diagnosis being delivered via an AI tool.

But if this were right it would prove too much. There are a great many diagnostic systems for which a doctor lacks any explanation. Doctors don't necessarily know how a particular lab test works, and thus why it produces the results that it does; nor exactly why an MRI machine produces the results it does. In general, a demand for explainability of this type would be far too onerous on doctors, and one rarely met. We generally allow that a doctor may not have an explanation for why a diagnostic system works, so long as they have sufficient confidence that it does in fact work.

The same seems true for AI tools. So long as a doctor has a high degree of confidence in the accuracy of the tool, we can assume that the doctor is acting in good faith when they use it. This, in turn, can underwrite the mental states implicated in accounts of strong trust. If a doctor acts in good faith by using the best diagnostic tools available, then we can be confident that they are acting with goodwill, or are responsive to the patient's reliance on them, or acting in line with a normative commitment of the right kind, and so on.

4.2 Objection Two

I have argued that explainability is not necessary for trust in AI that matters. Individuals can appropriately rely on AI without any explanation of the outputs that an AI produces.

Along the way, however, I noted that explainability might well *support* trust in AI. For instance, it is plausible that understanding why an AI system works will foster reliance on that system. Since reliance is a necessary condition on the seven notions of trust considered here, explainability thereby fosters trust in each sense. One might argue, however, that this is all anyone ever has in mind when it comes to linking explainability and trust. The idea that explainability is necessary for trust is a red herring.

There are indeed authors who focus on the idea that explainability *fosters* trust in AI rather than being necessary for it (see Kästner et al., 2021 for an overview and Blanco, 2022 for discussion). There are also empirical studies providing a strong indication that explainability does indeed foster trust in AI (see, for instance, Leichtmann et al., 2023).[9] But, as noted in §1, there are others who take explainability to be necessary for trust. The arguments in this paper can be seen as a corrective to this stronger trend in the study of AI.

4.3 Objection Three

Third and final objection: I have argued that it is not appropriate to have trust in AI of the strong type already discussed. This is because, roughly, AI systems don't have the right kinds of mental properties or mental states for any strong notion of trust to be appropriate. One might concede the point but urge nonetheless that it is better if individuals do in fact have strong trust of the relevant type, even if that trust is not appropriate. The idea here is that it is useful if individuals believe that AI systems act with goodwill toward them, even if current AI systems are not the kinds of things that have goodwill.

In other words, trust in AI in a strong sense is a useful fiction. What makes it useful? Well, if people are already inclined to believe albeit incorrectly that AI systems have mental properties or mental states, then it would perhaps make it more likely for people to be comfortable with the use of those systems if they believe AI is positively inclined toward them.

Perhaps that is right. But notice now that even if explainability is necessary for this type of strong trust, the types of explanations at issue will be strictly false. They will be explanations in terms of the mental properties of the relevant AI system, since as noted it is explanations of this type that seem to be important for stronger notions of trust. While this does leave some room for explainability to play a role, the space left has little to do with actual AI systems. For those systems don't have the relevant mental properties, so no correct explanation of the relevant kind can be given. We are left with making up fictional explanations to soothe the minds of those engaged in a useful fiction. This is a far cry from current explainability methods.

[9] We should be cautious in interpreting such studies. Such studies show that explainability fosters trust for a general notion of trust. What we need to do is break down trust by *type* and consider the relationship to explainability for each type.

5 Final Thoughts

I have considered seven types of trust, and have argued that explainability is either not necessary for trust, or that when it is, trust of the relevant kind is not appropriate for AI. As previously mentioned, the argument that I have presented focuses on local explainability. Does it generalise to global explainability? It depends on what global explainability is. On some approaches, global explainability is roughly a matter of aggregating a range of local explainability results into a more general model (Setzu et al., 2021). In this situation, it is unclear that global explainability would differ in kind from local explainability. That being so, we can expect the kind of argument I have presented here to generalise quite easily.

Whether the argument generalises for all notions of global explainability would require a more detailed look at the differences between global and local explainability. I lack the space to develop this point here. Notice, however, that none of the arguments presented above hang on the distinction between local and global explainability. That being so, it is likely the arguments will indeed generalise, and thus that neither local nor global explainability are necessary for trust in AI that matters. Determining whether this is in fact the case is a useful direction for future research.

It is tempting to view what I have said in this paper as an attack on explainability. Trust is often cited as one of the main motivations for explainability, and so if explainability is not after all needed for trust, then that seems to diminish the importance of explainability. Is my claim, then, that reliance or, at best, trust in some moderate sense is all we need for AI and that we should dispense with explainability altogether?

That's not the claim being made in this paper. The connection between explainability and trust is just one role cited for explainability. Explainability has been linked to a number of other important roles in the context of AI, including fairness (Orphanou et al., 2022), accountability (Doshi-Velez et al., 2019; Shin, 2021), and alignment with human values (Sanneman & Shah, 2023). It is thus likely that explainability is important for the ethical legitimacy and long-term public acceptance of AI systems, as well as for regulatory compliance, even if it is not necessary for trust.

Explainability can also be important in cases where one is using AI as a tool for scientific discovery (see, e.g., Wu et al., 2023), in part because it can be helpful for uncovering mechanisms or causal processes (for discussion, see Sullivan, 2022). This is important in healthcare, where such processes are important to discover (e.g., for drug discovery, see Jiménez-Luna et al., 2020). Indeed, explainability is particularly important in healthcare where biases or safety concerns with AI could have significant ethical implications (Amann et al., 2020; Combi et al., 2022). In these cases, explainability likely plays a large role both in ensuring procedural fairness and in ensuring alignment between the use of AI and the demanding ethical standards operative in a medical setting. In these cases, a demand for explainable AI systems is likely to be a crucial component of a comprehensive regulative framework for the responsible use of AI.

I am thus not recommending that we abandon explainability. Am I recommending that we take the focus away from explainability to some degree? The arguments presented here do not support even that conclusion. Rather, what the arguments may recommend is a reorientation of the discussion of trust and explainability. Rather than

trying to look at explainability exclusively through the lens of trust, perhaps we should focus on the benefits of explainability independently of trust, given that the relationship between explainability and trust is not as strong as some have previously thought.

To be clear: decoupling trust and explainability to a certain extent does not diminish the importance of explainability (or, indeed, trust). That's because, as noted above, explainability plays a number of important roles for AI. Decoupling explainability from trust does not undermine the capacity of explainability to play these roles. If anything it makes it clearer why explainability matters, and thus gives us a better sense of what to focus on when developing explainable AI systems.

Acknowledgements No acknowledgements.

Author Contributions All authors contributed equally to all aspects of the manuscript, including but not limited to concept, design, drafting and approving the final version. All authors agree to agree to be accountable for all aspects of the work.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of Data and Material Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable

Consent for Publication Not applicable.

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Aggarwal, N. (2021). The norms of algorithmic credit scoring. *The Cambridge Law Journal, 80*(1), 42–73.

Ali, S., Abuhmed, T., El-Sappagh, E., Muhammad, K. M., Alono-Moral, J., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What We Know and What is Left to Attain Trustworthy Artificial Intelligence. *Information Fusion, 99*, 101805.

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law, 30*, 1–17.

Alonso, F. (2016). Reasons for Reliance. *Ethics, 126*(2), 311–338.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education, 23*(1), 689.

Amann, J., Blasimme, A., Vayena, E., Dietmar F., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making 20*(310). https://doi.org/10.1186/s12911-020-01332-6

Baier, A. (1986). Trust and Antitrust. *Ethics, 96*(2), 231–260.

Beisbart, C., & Räz, T. (2022). Philosophy of science at sea: Clarifying theinterpretability of machine learning. *Philosophy Compass, 17*(6), e12830. https://doi.org/10.1111/phc3.12830

Belkin, M., Hsu, D., Siyuan, M., & Soumik, M. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Blanco, S. (2022). Trust and explainable AI: promises and limitations. *Proceedings of the ETHICOMP*, (pp. 246–257).

Buijsman, S. (2023). Over What Range Should Reliabilists Measure Reliability? *Erkenntnis, 89*, 2641–2661.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2022). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *Journal of Artificial Intelligence and Consciousness, 9*(1), 59–72.

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? *Boston Review* 1.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A Review of Trustworth and Explainable Artificial Intelligence (XAI). *IEEE ACCESS, 11*, 78994.

Choung, H., Prabu, D., & Ross, A. (2021). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human?Computer Interaction 23*(9), 1727–1739.

Colaner, N. (2022). Is explainable artificial intelligence intrinsically valuable? *AI & Society, 37*, 231–238.

Combi, C., Amico, B., Bellazzi, R., Holzinger, A., Moore, J. H., Zitnik, M., & Holmes, J. H. (2022). A manifesto on explainability for artificial intelligence in medicine. *Artificial Intelligence in Medicine, 133*, 102423.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, *87*(4), 568–589. https://doi.org/10.1086/709729

Dam, H. K., Tran, T., & Ghose, A. (2018). Explainable software analytics. *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, (pp. 53–56).

Donnelly, L. (2017). Forget your GP, robots will 'soon be able to diagnose more accurately than almost any doctor' *The Telegraph*.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2019). Accountability of AI Under the Law: The Role of Explanation. arXiv:1711.01134

Duede, E. (2022). Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese, 200*(6), 491.

Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science, 90*(5), 1089–1099.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis oftrust in medical AI. *Journal of Medical Ethics, 47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Dressel, J. & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.

Ferrario, A., & Loi, M. (2022). *How Explainability Contributes to Trust in AI FAccT '22, June 21–24*. Seoul: Republic of Korea.

Freiesleben, T., & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese, 202*(4), 109. https://doi.org/10.1007/s11229-023-04334-9

Frieman, O. (2023). Making sense of the conceptual nonsense 'trustworthy AI'. *AI and Ethics, 3*, 1351–1360.

Fox, M., Long, D., & Magazzeni, D. (2017). Explainable Planning. *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, (pp. 24–30).

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: an overview of interpretability in machine learning. *IEEE 5th International Conference on Data Science and Advanced Analytics DSAA*, (pp. 80–89).

Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019). Explainable AI in Industry. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (pp. 3203–3204).

Goldberg, S. C. (2020). Trust and Reliance. In J. Simon (Ed.), *The Routledge Handbook of Trust and Philosophy* (pp. 97–108). New York: Routledge.

Grote, T., Genin, K., & Sullivan, S. (2024). Reliability in Machine Learning. *Philosophy Compass*, e12974. https://doi.org/10.1111/phc3.12974.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ... & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, *316*(22), 2402–2410.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, Ben Hadj, & A., Thomas, L., Enk, A., & Ulhmann, L. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology, 29*(8), 1836–1842.

Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs, 48*(1), 1–20.

Holton, R. (1994). Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy, 72*(1), 63–76.

Hong, Y., Lian, J., Xu, L., Wang, Y., Freeman, L. J., & Deng, X. (2023). Statistical perspectives on reliability of artificial intelligence systems. *Quality Engineering, 35*(1), 56–78.

Jiménez-Luna, J., Grisoni, F., & Shneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence, 2*, 573–584.

Jones, K. (1996). Trust as an Affective Attitude. *Ethics, 107*(1), 4–25.

Jones, K. (2012). Trustworthiness. *Ethics, 123*(1), 61–85.

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. arXiv:2108.05379v2

Kelly, S., Kaye, S., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics, 77*, 101925.

Kuncel, N. R., Ones, D. S., & Klieger, D. M. (2014). In Hiring, Algorithms Beat Instinct. *Harvard Business Review*.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650.

Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. *Intelligent Computing: Proceedings of the Computing Conference*, (pp. 1269–1292).

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

McLeod, C. (2002). *Self-Trust and Reproductive Autonomy*. Cambridge, MA: MIT Press.

Nahmias, E., Allen, C. H., & Loveall, B. (2020). When do robots have free will? Exploring the relationships between (attributions of) consciousness and free will. In B. Feltz, M. Missal, & A. C. Sims (Eds.), *Free will, causality, and neuroscience* (pp. 57–80). Brill Publishers.

Nickel, P. (2013). Trust in Technological Systems. In M. de Vries, S. Hansson, & A. Meijers (Eds.), *Norms in Technology* (pp. 223–37). Dordrecht: Springer.

Nickel, P. (2022). Trust in Medical Artificial Intelligence: A Discretionary Account. *Ethics and Information Technology, 24*, 7.

Nguyen, C. T. (2022). Trust as an Unquestioning Attitude. In J. Hawthorne, J. Chung, & T. Gendler (Eds.), *Oxford Studies in Epistemology* (Vol. 7, pp. 214–244). Oxford: Oxford University Press.

Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., Shulner Tal, A., Hartman, A., & Kuflik, T. (2022). Mitigating Bias in Algorithmic Systems? A Fish-eye View. *ACM Computing Surveys, 55*(5), 1–37.

Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology, 13*(1), 53–64.

Rachman, S. (2010). Betrayal: A psychological analysis. *Behaviour Research and Therapy, 48*, 304–311.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778.

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics, 26*, 2749–2767.

Sanneman, L., & Shah, J. (2023). Transparent Value Alignment. *HRI '23: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 557–560).

Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., & Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. *Frontiers in Computer Science, 5*,. https://doi.org/10.3389/fcomp.2023.1151150

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

Schoeffer, J., De-Arteaga, M., Kühl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems 836*, 1–18.

Seddik, B., Ahlem, D., & Hocine, C. (2022). An Explainable Self-Labeling Grey-Box Model. *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 1–7. https://doi.org/10.1109/PAIS56586.2022.9946912

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence, 294*, 103457.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*.

Stout, R. (2022). Betrayal, Trust and Loyalty. *International Journal of Philosophical Studies, 30*(3), 339–356.

Sullivan, Y., de Bourmont, M., & Dunaway, M. (2022). Appraisals of harms and injustice trigger an eerie feeling that decreases trust in artificial intelligence systems. *Annals of Operation Research, 308*, 525–548.

Sullivan, E. (2022). Understanding from Machine Learning Models. *British Journal for the Philosophy of Science, 73*(1).

Sullivan, E. (2023). Do Machine Learning Models Represent Their Targets? *Philosophy of Science, 91*(5), 1445–1455.

Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*(2), 243–257. https://doi.org/10.1007/s11023-010-9201-3

Taddeo, M. (2011). Defining trust and e-trust. *International Journal of Technology and Human Interaction, 5*, 23–35. https://doi.org/10.4018/jthi.2009040102

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology, 34*, 1607–1622.

Vandenburgh, J. (2023). Machine Learning and Knowledge: Why Robustness Matters. arXiv:2310.19819

Yang, S., Krause, N. M., Bao, L., Calice, M. N., Newman, T. P., Scheufele, D. A., Xenos, M. A., & Brossard, D. (2023). In AI We Trust: The Interplay of Media Use, Political Ideology, and Trust in Shaping Emerging AI Attitudes. *Journalism & Mass Communication Quarterly. Online First:* https://doi.org/10.1177/10776990231190868

Wachter, S., Mittlestadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Wu, Z., Chen, J., Li, Y., Yafeng, D., Zhao, H., Chang-Yu, H., & Hou, T. (2023). From Black Boxes to Actionable Insights: A Perspective on Explainable Artificial Intelligence for Scientific Discovery. *Journal of Chemical Information and Modeling, 63*(24), 7617–7627.