How Explainability Contributes to Trust in AI

Andrea Ferrario*
aferrario@ethz.ch
ETH Zurich
Switzerland

Michele Loi
michele.loi@polimi.it
Politecnico di Milano
Italy

ABSTRACT

We provide a philosophical explanation of the relation between artificial intelligence (AI) explainability and trust in AI, providing a case for expressions, such as "explainability fosters trust in AI," that commonly appear in the literature. This explanation relates the justification of the trustworthiness of an AI with the need to monitor it during its use. We discuss the latter by referencing an account of trust, called "trust as anti-monitoring," that different authors contributed developing. We focus our analysis on the case of medical AI systems, noting that our proposal is compatible with internalist and externalist justifications of trustworthiness of medical AI and recent accounts of warranted contractual trust. We propose that "explainability fosters trust in AI" if and only if it fosters justified and warranted paradigmatic trust in AI, i.e., trust in the presence of the justified belief that the AI is trustworthy, which, in turn, causally contributes to rely on the AI in the absence of monitoring. We argue that our proposed approach can intercept the complexity of the interactions between physicians and medical AI systems in clinical practice, as it can distinguish between cases where humans hold different beliefs on the trustworthiness of the medical AI and exercise varying degrees of monitoring on them. Finally, we apply our account to user's trust in AI, where, we argue, explainability does not contribute to trust. By contrast, when considering public trust in AI *as used by a human*, we argue, it is possible for explainability to contribute to trust. Our account can explain the apparent paradox that in order to trust AI, we must trust AI users not to trust AI completely. Summing up, we can explain how explainability contributes to justified trust in AI, without leaving a reliabilist framework, but only by redefining the trusted entity as an AI-user dyad.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**; • **Applied computing → Sociology**; • **Social and professional topics → Computing / technology policy**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

artificial intelligence, explainable artificial intelligence, trust, healthcare, trustworthiness, ethics of artificial intelligence

ACM Reference Format:
Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533202

1 INTRODUCTION

Explainability is the desideratum of artificial intelligence (AI) that would allow humans to understand the inner logic of these systems and the rationale behind the generation of their outcomes. It is either a requirement satisfied "by design" by a restricted class of algorithms, called "interpretable models," or it is strived for by means of methods approximating the AI globally or locally, at each AI outcome [29].[1] The interdisciplinary literature on AI suggests that explainability of AI may engender, increase, or generate trust [8, 29, 33]. The possibility of engendering (levels of) trust in AI using explainability becomes particularly relevant for the case of medical AI systems (medical AIs). This is due to 1) their use in high-stakes and high impact scenarios [22], 2) the recent rise of their (contested) epistemic authority as the result of high-profile works in a number of applications [22], 3) the higher cost of alternatives for decision-making [45], and 4) the ethical challenges arising from their use in clinical decision-making processes [5, 7, 14]. Without a clear definition of the relation between "explainability" and "trust," claims such as "explainability fosters trust in AI" and cognates remain unclear. This affects explainable AI and its far-reaching promises, i.e., "to audit, to validate, to discover" [44],[2] thus contributing to the low rate of acceptance and deployment of medical AI in clinical practice [41].

This paper takes on this challenge in a two-step approach. In the first, conceptual, step, we discuss whether and how explainability fosters trust in AI. To do so, we discuss and combine three conceptual building blocks: 1) *justification*, 2) *causality*, and 3) *monitoring*. *Justification* refers to a recent discussion on the explainability of medical AIs suggesting that the epistemic justification of the beliefs on their trustworthiness is necessary to trust these systems [14]. Therefore, only by providing appropriate warrants of the reliability of the medical AI the trustworthiness of its outcomes can be

*Both authors contributed equally to this research.

[1] In this paper, we use the term "explainability" to refer to the provision of explanations of machine learning outcomes (local explanations), or its "inner working" (global explanations) [31, 35]. Explanations can be model-agnostic or model-specific and are generated by algorithmic procedures that we call "explainability methods."

[2] In the case of counterfactual explanations [43] of AI outcomes, one can also recommend actionable strategies to achieve alternative and more favourable outcomes (as opposed to, for example, being denied credit by an algorithmic credit lending AI system) [42].

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533202

justified [14]. These warrants[3] can be realized by considering, for example, "computational reliabilism," as an answer to the opacity of black box algorithms used in medical AIs [13, 14, 23]. The question is whether the methods of explainable AI may be regarded, in the computational reliabilism framework, as examples of epistemic warrants, i.e., realiability indicators [14].[4] *Causality* refers to the epistemic warrant provided by the causal and counterfactual dependence between the belief in the trustworthiness of medical AI and the actual trustworthiness of AI [24]. Finally, *monitoring* refers to the intuition that if X trusts Y, X is willing to rely on Y without (or with only a little, in relative terms) monitoring or supervision of Y [3, 11, 18, 27, 30, 40]. This account of trust, "trust as anti-monitoring," suggests that where a reliance relation subsists with positive expectation and no (or little) monitoring, there is (high) trust. Some authors also introduce quantitative models to measure levels of trust [30, 40]. In a sense, trust is simply reliance in the absence of monitoring, as opposed to a common approach in the philosophical literature on reliance and trust that tends to moralize trust [32]. Toward that end, we discuss computational reliabilism, warranted trust, and we introduce three forms of trust from the "trust as anti-monitoring" account: simple, reflective, and paradigmatic [18, 30].

Combining the three conceptual blocks, we argue that claims, such as "explainability fosters trust in AI," should consider the epistemic justification of the trustworthiness of the system, the perceived need of stakeholders to monitor an AI, and "connect" the trustworthiness of the AI with the activity of monitoring appropriately. Therefore, we claim that "explainability fosters trust in AI" iff explainability supports the justified belief on the trustworthiness of an AI and because of this belief it leads to a decrease of monitoring activities. To support means two distinct relations occurring concurrently: first, a relation of justification: it is in the light of explainability that the belief in trustworthiness is justified. Second, a relation of causality: explainability is a property that makes AI trustworthy and is also part of what causes the belief that the AI is trustworthy, which, in turn, causes lower monitoring levels. The consequent of the conditional is equivalent to requiring that explainability fosters justified and warranted paradigmatic trust [18, 24, 30].

In the second, practical step, we explain our account by introducing the example of a physician interacting with a medical AI in clinical practice. To do so, we show that different forms of trust in medical AI emerge from an analysis that considers four types of trust, resulting from two different possible levels of justification and two possible levels of monitoring invested in a reliance relation. Therefore, our account to explainability and trust in AI allows us to distinguish different cases of real-world human-AI interactions and inform future human-AI interaction studies on explainability and trust. Moreover, we describe some empirical conditions that enable relations characterized by justified and warranted paradigmatic trust between physicians and medical AIs. We apply our account of the relation between explainability and trust to 1) a user's trust in AI, and 2) the public trust in what we call the "AI-user dyad," in the paradigm of augmented (as opposed to artificial) intelligence. If our analysis is correct, the future step in analyzing the explainability of AI and trust should be: given a specific context, i.e., a medical workflow, to identify the properties of explainability methods that justify a belief on the trustworthiness of the AI, and of the AI-user dyad, which, in turn, are likely to cause a justified measurable reduction of its monitoring. Moreover, we provide a sketch for an explanation (based on *a priori* considerations) of how, in the case of the AI-user dyad, explainability may, indeed, contribute to trust.

The paper is organized as follows. In Section 2, we present an overview of explainability and trust in AI. In Section 3, we introduce the building block *justification* by discussing computational reliabilism in medical AI [13]. In Section 4, we discuss the building block *causality* by commenting on the account of "warranted trust" by Jacovi et al. [24]. In Section 5, we introduce the building block *monitoring*, providing an overview of the main accounts of "trust as anti-monitoring." In Section 6, we introduce our account of "explainability fosters trust in AI," we discuss its feasibility by relating the different types of justifications of trustworthiness of medical AI to the account of trust as anti-monitoring, and we distinguish the trust of the AI by its user from the trust of the public in the AI-user dyad. In Section 7, we present our considerations against thinking it plausible that explainability will contribute to *the user's* trust in AI, to the extent that this is justified. Finally, in Section 8, we discuss a distinct trust relation, where the trusted entity is a combination of AI and human (the AI-user dyad) and the trustor is the public (e.g., patients, as opposed to physicians). We explain how AI explainability may possibly contribute to *this* relation of trust. We then present our conclusions.

2 EXPLAINABILITY AND TRUST IN AI

Explainability of AI is an elusive concept. Originally considered to be an answer to the problem of under-specification of machine learning models embedded in AI systems [29], it consists of providing users with 1) understandable information on the "inner working" of the model, and 2) explanations of the model outcomes [29]. In the explainable AI research domain, scholars defend different positions around interpretability, providing a different rationale for its relation with concepts, such as understanding, explainability, and transparency [12, 16, 29, 33, 34, 47]. Independently of the positions taken on these relations, claims about the explainability of AI refer to a standard, namely "explainable AI," comprised of algorithms that are explainable by design. The structure of these models, also called "interpretable models," is simple enough to be fully understood through a sequence of cognitively accessible deductive arguments.[5] Therefore, the aim of explainability is to approximate the degree of explainability of "interpretable models" by using methods to improve the understanding of a "black box" model logic and its outcomes. In the literature, it is often stated that the explainability of an AI may support the development of users' trust in the machine learning models it uses and increase

[3] We will clarify our use of the term "warrant" in Section 4: see footnote 10.

[4] As it will become clearer in section 7 and 8, our answer is neither affirmative nor negative in an absolute sense. For we argue that explainability can hardly be seen as a reliability indicator of AI *in itself*, but is a plausible one for the AI-user dyad.

[5] Examples of interpretable algorithms are decision trees, rule-based, and linear models [31]. However, the property of being inherently interpretable is possibly a function of the model complexity and the number of input features: a decision tree with a thousand nodes or a linear regression in hundreds of variables and their interactions are arguably not more interpretable than other models [31].

their acceptance rate [16, 19, 25]. While scholars describe this development using different expressions, all of them suggest that an appropriate use of explainability methods would foster trust in the machine learning model and, by synecdoche, in the AI system. As remarked by Bjerring and Busch, "if we can explain the reasons behind a certain AI prediction, then it is more likely that people will trust the AI system and act on the prediction" [7]. To this end, when considering algorithms in medical AI systems, Binns et al. remark that "the provision of explanations can affect levels of trust and acceptance of algorithmic decisions" [6]. In the literature, the verbs used to describe the relationship between the explainability of AI and trust are "to engender" [29], "to increase" or "to generate" [33], "to gain" [8], and "to build" [31], among others. However, we note that the dynamics between the explainability of AI and the trust of the human agents interacting with those systems is far from being clarified. This is primarily due to the lack of a precise definition of both the explainability of AI and trust in AI. This affects, *a fortiori*, the dynamics of their relation, as expressed by the verb "to engender" and its cognates.

To model the relation between explainability and trust in AI, inspired by recent discussions in the literature on the philosophy of AI and human-AI interactions [14, 18, 24, 40], we introduce and discuss three conceptual building blocks: 1) *justification*, 2) *causality*, and 3) *monitoring*. We aim to show how these building blocks support a philosophical account of how explainability fosters trust in AI that can inform future studies on human-AI interactions. We focus our analysis on medical AIs, due to their increasing importance in healthcare applications and the relevance of discussions on the epistemic quality of their outcomes. In fact, high-profile scientific contributions show that AIs reach or surpass the performance of human experts in selected clinical tasks, such as in the classification of medical scans and images in radiology, oncology, and dermatology [10, 17, 21, 41]. *De facto*, this body of literature contributes to the perception of medical AIs as a trustworthy guide to clinical shared decision-making. As medical AIs are typically meant to operate in high-stakes scenarios, it is necessary to carefully validate the epistemic quality of medical AI outcomes. In fact, alternative scenarios, such as contacting other healthcare experts, scheduling further medical examinations to generate additional evidence, may be unfeasible or too expensive for both the healthcare system and patients [45].

3 JUSTIFICATION AND TRUST IN AI

We introduce the first conceptual block of our approach, i.e., *justification*, by discussing Durán and Jongsma's recent analysis on the use of black box algorithms in medical AIs [14]. The authors state that, when considering interactions with medical AIs, the epistemology of algorithms precedes the study of the ethics of algorithms [14]. In fact, they argue that the main epistemic obstacle to the trustworthiness of medical AIs (or, more precisely, of their outcomes) is represented by algorithmic opacity [23], i.e., the impossibility for an agent "to have access to and be able to survey all of the relevant elements of the justification" [13], due to the very nature of the algorithm. As a result, algorithmic opacity makes "it impossible to ground the reliability of the algorithm and, consequently, on whether researchers, physicians and patients can trust the results of such systems" [14]. Therefore, Durán and Jongsma propose the theory of computational reliabilism (CR) [13] as a solution to the problem of epistemic opacity of medical AIs. CR states that "researchers are justified in believing the results of AI systems because there is a reliable process (ie, the algorithm) that yields, most of the time, trustworthy results" [14]. In other words, CR provides epistemic justifications for the belief that an opaque algorithm is reliable and its outcomes are trustworthy through processes that are exogenous to the algorithm, without relying on the use of explainers (such as in the case of explainable AI), or the necessity to renounce to the use of opaque algorithms *in toto* [14]. Embracing CR in an epistemology of medical AIs we can conceive a *de facto* successful (meaning, truth-conducive, and error-reducing) interaction with the system as what it makes a belief in its trustworthiness justified. Thus we identify, at least *prima-facie*, the concept of AI trustworthiness with that of its reliability.[6]

If the interaction between a user and an AI is a reliable process for judging the reliability of the AI, the beliefs about the trustworthiness of the AI are justified. Therefore, the beliefs of, e.g., a physician that are produced by such interaction will be justified. A candidate reliable process is an interaction in which a physician may compare the congruence between the diagnostic outputs of an AI and the diagnosis that he would have produced without the assistance of the AI.[7]

There are two open challenges that arise from embracing CR as a framework to justify the beliefs on the trustworthiness of a medical AI. The first deals with adapting the theory of CR to the case of black box algorithms in a medical AI, and the second with defining the role of the explainability of an AI in the context of CR. In fact, Durán and Jongsma notice that CR was originally introduced in the case of computer simulations [13]. Therefore, they argue that the reliability indicators of black box algorithms in medical AIs must be adapted to the specificities of these algorithms, and they also emphasize the need to introduce new ones [14]. However, they both state that "more transparency[8] in algorithms is not always necessary" [14] and that "transparency [...] by itself is necessary, although not sufficient for entrenching the reliability of black box algorithms and the overall trustworthiness of their results" [14]. Confusion results, but we think that within the reliabilism framework they offer, the first claim (that transparency is not necessary) is more justified (unless reliability is not intrinsic to the AI, but a property of what we shall label the "AI-user dyad," as we argue in Section 8).

[6] Although the label "computational reliabilism" uses the word "reliablism," it does not presuppose reliabilism as a theory about justification in its usual connotation, i.e., as externalist reliabilism. Internalists can all think that *beliefs* about the reliability of a belief forming method can provide justification for believing the outputs of that method. What characterizes an externalist position is for us relying on mere, untested reliability, which is not what [13] advocate. The idea that we can be justified in believing the outputs of systems *that we have tested for* reliability is compatible with an internalist theory stating that evidence of reliability is good for justification. So reliabilism about justification is just orthogonal to the issue of whether (externalist) reliability is adequate for trust, *pace* [13]. For one can be an internalist and agree that one can form justified beliefs on the basis of reliable methods, if one knows they are reliable, see [15, 39].

[7] For the importance of error-identifying and correcting processes involving humans, see [1]. Alvarado is skeptical about the human ability to correct (opaque) AI.

[8] We interpret transparency to be a form of explanainability in this context, because it is "a process that informs the inner workings of black box algorithms" [14].

4 CAUSALITY AND TRUST IN AI

We move on introducing the second building block of our approach: *causality*. Recently, Jacovi et al. have introduced an account of trust called "contractual trust," that aims at explaining the concept of trust in human-AI interactions [24]. Contractual trust relies on contracts, i.e., explicit functionalities (requirements) of an AI that may be deemed useful in a given context [24]. These contracts may follow from key requirements for trustworthy AI, such as accuracy, transparency, and non-discrimination, and they are supported by standardized documentation that can be used to increase trust in the contract itself [24]. In this context, as an AI "is trustworthy to some contract if it is capable of maintaining this contract" [24], and contractual trust is "to believe that a set of contracts will be upheld" [24],[9] then contractual trust is a belief in the trustworthiness (with respect to a contract) of an AI. On the other hand, trustworthiness of an AI is an objective property of the system that other authors refer to as "actual trustworthiness" [28, 38].

Jacovi et al. further investigate the relation between trustworthiness of an AI and trust in it by stating that only the forms of "warranted trust," that is trust non-accidentally caused by the trustworthiness of the AI, is ethically desirable.[10] Warranted trust is defined "via a causal (interventionist) relationship with trustworthiness" [24], which means that warranted trust is trust caused by the trustworthiness of the AI.[11] By definition, warranted trust can be tested in empirical studies where trustworthiness is manipulated across different experimental conditions and trust is quantified in (binary or continuous) levels [24]. However, if trust and trustworthiness both come in degrees, we observe that a causality claim, such as the one supporting warranted trust, does not justify a specific level of trust. In general, the level of trust of a human interacting with a medical AI can be disproportionate with respect to the level of trustworthiness of the AI, leading to cases of distrust or overtrust [28].[12] Jacovi et al. state that the addition of the explanation enables the user to better anticipate "whether the model's decision is correct or not for given inputs, [...] compared to the model without any explanation" [24]. Note that the ability of a user to evaluate correctness does not make the algorithm more accurate, robust, and reliable in itself; it can only, at best, make the use of the AI by the human more accurate, robust, and reliable. So, it is no longer clear what the object of trust is: whether it is trust of the user in the AI, or trust by the public of the AI as used by a specific user.

The above discussions on justification, causality and trust in AI bring us to the golden standard for trust in human-AI interactions: ideally, we would want trust in AI to be justified and warranted. In that case, we would say that the trustor holds a justified belief that the AI will uphold a set of contracts, and this belief is affected by the manipulation of the capability of the AI to maintain these contracts [24].[13] In what follows, we shall refer to trust that is justified and warranted as JW-trust (see Figure 1).[14]

Finally, we note that that discussions on justification (in its "calibrated with trustworthiness" form [28]), causality and trust in AI rely on the existence and measurement of levels of trust. However, neither Durán and Jongsma nor Jacovi et al. address the how and when of trust level existence and measurement,[15] and how to relate the measurement of trust in AI to the actual use of explainability methods. We will tackle on this challenges in the next sections.

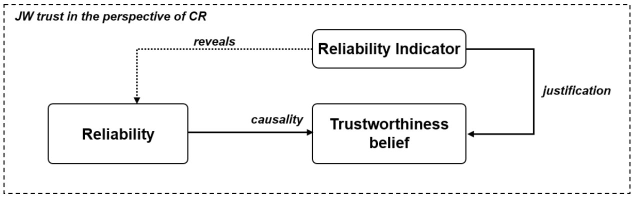

Figure 1: The concept of JW-trust. We bring together the concept of justification from CR and that of causality from warranted trust into a single model of the relations between reliability, reliability indicators, and beliefs in trustworthiness. This figure represents the resulting logical framework, which combines the two theories.

[9] We interpret the capacity of the AI to maintain relevant functionalities (e.g., being accurate), that is Jacovi et al.'s trustworthiness, as a form of reliability of the system. This said, one could try introducing a contract that revolves around explainability of the AI, but clearly, the argument that explainability may contribute to trustworthiness simply because it is included in a contract is question begging in this context.

[10] This view about the ethical desirability of trust is arguably too extreme. There could be specific occasions in which trust that is not warranted, or not justified, is beneficial to all the affected parties. Our use of "warrant" follows the definition of warranted trust provided in [24]. Notice that this is different from its usage within analytic philosophy, for example by [37, 46]. In [24] "warrant" does not mean whatever property distinguishes true belief and knowledge, but a specific property of counterfactual and causal dependence between the property "trustworthiness" and the belief in trustworthiness. These two meanings are connected. The causal connection between trustworthiness and the (true) belief in trustworthiness arguably is also a "warrant" in the philosophically established sense, because it discriminates between those beliefs in trustworthiness that merely happen to be true, e.g., Gettier cases [20], from those that are true *because* the AI is trustworthy.

[11] In [24] the cause of trust is either a match between user priors and observable AI decisions or a positive evaluation by a trustworthy evaluation method (the latter corresponds to the computational reliability idea). Our account is more general and abstract: trustworthiness causes trust, if the mental representation of that property – a belief in trustworthiness, itself caused by trustworthiness – is what causally explains the reduction of monitoring (see Section 5). As an alternative to warranted trust, one could trust an AI without a causal relationship with trustworthiness, or not trust it in the presence of trustworthiness: the former is "unwarranted trust," or an ethically unacceptable form of trust in AI systems that "should be explicitly evaluated against, and avoided or otherwise minimized" [24], the latter is a failure to develop warranted trust.

[12] In the literature on human-machine interactions, the appropriateness of trust levels, as related to those of trustworthiness, is usually encoded by the expression "calibrated with trustworthiness" [24, 28]. This expression, originally introduced by Lee and See [28] states that trust "matches [AI] capabilities, leading to an appropriate use" [28]. Therefore, we interpret "calibrated with trustworthiness" as "global" justified trust: all levels of trust are appropriate, for all levels of trustworthiness of the AI [28]. However, calibrated trust does not need to be warranted.

[13] An example of warranted but not justified trust in an AI: a physician may stop validating the outcomes of a medical AI, therefore manifesting overtrust, because the system he operates on has a new functionality that allows users to access summary statistics of the data used for its training. We will return on this point in Section 7.

[14] We believe that a discussion of the case of trust that is justified but *not* warranted would be of little practical interest to devote extensive attention to it. For the most epistemologically conscious reader, it can be thought as analogous to a Gettier case [20], where the belief in the AI being trustworthy happens to be true and justified, but not *because* the AI is trustworthy. (Counterfactually: had the AI *not* been trustworthy, the trustor would have still believed it was.)

[15] Jacovi et al. mention the possibility to measure trust on a sliding scale [24]. However, they argue that a single scale could not be appropriate to measure trust, as this is a multi-dimensional construct. We agree on this point, if the single scale simply captures self-assessed levels of perceived trust [26]. We discuss alternative quantifications of trust in Section 5.

5 MONITORING AND TRUST IN AI

We introduce the last block of our approach, i.e., *monitoring*, by discussing those activities that a trustor may engage in after making the decision to rely on another agent (i.e., the trustee) and their relation to trust in AI. We label these activities simply as "monitoring."[16] Trust understood as essentially antithetical to monitoring ("trust as anti-monitoring"), or the statement that trusting means not controlling the actions of whom is deemed worthy of trust, has been discussed by different authors in the philosophical literature [3, 4, 11, 27]. Trust as anti-monitoring pertains also to trusting relations involving both human and artificial agents (AA). For example, in the context of e-trust, Taddeo discusses the relation between the trustworthiness of an AA and its supervision (i.e., monitoring) [40]. The absence of supervision is justified by the trustee's trustworthiness, which leads to the minimisation of the trustor's effort and commitment [40]. Moreover, considering e-trust in AA-AA interactions, Taddeo states that the level of resources saved by the trustor is inversely related to its level of e-trust [40].

More recently, Ferrario et al. further explored trust as antithetic to monitoring by introducing its three different forms, called "simple," "reflective," and "paradigmatic" [18]. They state that simple trust is a disposition to rely without monitoring in the absence of beliefs about trustworthiness. Reflective trust is the mere having of beliefs about trustworthiness. Finally, paradigmatic trust is when both co-exist: X trusts Y paradigmatically iff X holds a belief on Y's trustworthiness and relies on Y without monitoring [18]. Ferrario et al.'s account is applicable to interpersonal and human-AI interactions. In the case of paradigmatic trust in AI, the trustor is disposed to rely on the AI without monitoring it and deems the AI to be trustworthy [18]. Most famously, an account of trust as anti-monitoring has been (implicitly) assumed by Baroness O' Neill in her BBC Reith Lectures [36]. O' Neill explained the dangers for trust implicit in attempts to rebuild public trust by designing tighter and tighter mechanisms of institutional control. For control expresses lack of trust.

We can combine the aforementioned account of JW-trust with the account of trust as anti-monitoring *a priori*, as follows. We state that JW paradigmatic trust is a monitoring-avoiding relation where the trustor holds a warranted and justified belief on the trustworthiness of the trustee (a true justified belief caused by trustworthiness), which causes the trustor to avoid monitoring of the trustee, or to reduce it to a specific degree, and justifiably so. Given this definition, Ferrario et al.'s simple (defective) trust is unjustified and unwarranted from an internalist point of view [18].[17] It can be the result of contingent (e.g., lack of time) or structural (e.g., the impossibility to entertain beliefs on trustworthiness) causes. However, paradigmatic trust can be justified and warranted. In this case, the absence of monitoring results from a justified belief on the trustworthiness of the trusted entity, where the belief, in turn, is a reasonable response to the features making the entity trustworthy.[18]

6 HOW DOES EXPLAINABILITY FOSTER TRUST IN AI?

We wrap-up the discussions from Section 3, 4 and 5, and we introduce our philosophical account of how explainability fosters trust in AI. We apply the same account to analyze user's trust in AI and the public trust in the AI-user dyad, that is to say, the AI *as used* by a specific user.

Let us consider the illustrative example of a physician interacting with a medical AI in clinical practice.[19] The key point is the link between the nature (JW or not, or even absent) of the beliefs on the trustworthiness of the medical AI and the exercise of monitoring throughout the reliance relation. For the sake of simplicity, we only assume two possible levels of monitoring, low and high, corresponding to trust and its absence. We also assume a simplification of the different normative qualities of trust, by assuming that beliefs in the trustworthiness are either, simultaneously, justified and warranted, or that they lack those three qualities altogether (or are entirely absent). This results in four combinations that provide a simplified spectrum of different real-world scenarios where humans rely on medical AI. These considerations are summarized in Table 1.[20]

	Low monitoring (high trust)	**High monitoring (low trust)**
(JW) belief on trustworthiness	(JW) paradigmatic (high) trust. *(Technology-savvy and trained physician)*	Mere reflective trust. *(Critical/careful/zealous physician)*
No (JW) belief on trustworthiness	Mere simple (high) trust or unwarranted paradigmatic (high) trust. *(Naïve/technology-enthusiast physician)*	Very low or no trust. *(Skeptical physician)*

Table 1: Four different types of trust of a physician in AI. The four types result from the intersection of the two independent dimensions of trust, namely belief in trustworthiness and monitoring. For the sake of simplicity, here we consider only two types of trustworthiness beliefs (the AI is either believed trustworthy or not) and two levels of monitoring (low and high).

[16] We follow Castelfranchi and Falcone to define monitoring, i.e., the activity "aimed at ascertaining whether another action has been successfully executed" [9]. When considering trusting relations, monitoring is subsequent to the decision to trust and the delegation to an agent (the trustee) of a set of actions to reach a goal of interest for the trustor [9]. We will discuss the activities of monitoring in the case of physician-medical AI interactions in Section 6.

[17] If my trustee is reliable, then a reliabilist will say I am justified in believing what the trustee tells me, regardless of whether I have any beliefs about the trustee.

[18] To be precise, Ferrario et al. do not discuss degrees of trust that are implicit in the idea of calibrated trust. A gradualist analysis of trust is provided by Taddeo [40]. This account has been combined with the one in [18], leading to the theory exposed in [30].

[19] For example, a medical AI may classify medical images, predict the likelihood of hospital readmission, or selected patient outcomes, such as stroke and sepsis, in a given time frame.

[20] The information in Table 1 considers the absence vs. presence of monitoring as introduced by Ferrario et al., for the sake of simplicity [18]. It can be straightforwardly modified including degrees of monitoring, such as in Loi et al.'s account [30].

With reference to Table 1, the use of the "Technology-savvy and trained physician" would typically involve no (or very little) monitoring as he used the medical AI many times in the past, identifying its error patterns and discussing its internal logic in-depth with the data engineers responsible for its design. These activities allowed him to form a well-posed belief on the trustworthiness of the medical AI. If the absence of monitoring results from this well-posed belief, then he exercises JW paradigmatic trust in the AI. However, a "Critical/careful/zealous physician" would exercise (a high level of) monitoring. He might believe that the medical AI is useful but its output still too unreliable to be blindly trusted. The physician believes that the AI is trustworthy (it achieves what it promises) but only if it is supervised to avoid obvious errors, which a skilled physician ought to recognize. A "Naïve/technology-enthusiast physician" could rely on the medical AI exercising no (or very little) monitoring on it, due to having formed an unjustified belief on its trustworthiness based, for example, on the superficial testimony of peers or the high level of usability of its interfaces. This would be an instance of simple trust [18], that is not JW. Finally, the "Skeptical physician" may justifiably hold no belief on the medical AI being trustworthy, as he started in his current position a few days prior and had never used the AI tool before. This would be a case of where trust is absent, and could also be considered a limiting case of JW-trust where calibrated and warranted trust is justifiably very low. Therefore, the exercise of monitoring (even very high levels of it) would follow as a tentative step to assess whether the reliance relation goal had been achieved and to justify the triggering of alternative decision-making processes (e.g., consulting other physicians).

In summary, considering the case of a medical AI and inspired by the possible scenarios described in Table 1, we say that **explainability fosters trust in AI iff explainability contributes to paradigmatic JW-trust in reliance relations with the medical AI**. This is equivalent to stating that explainability supports the justified belief on the trustworthiness of the medical AI, and because of this belief it leads to the reduction of the degree of monitoring invested in the reliance relation.[21] Notice that "support" here refers to a normative relation endowed with causal power. Explainability not only provides the belief with the normative quality of being justified, but also the causal explanation of that belief, as required by the idea of warranted trust. Therefore, the higher the degree of monitoring, the lower the level of paradigmatic JW-trust [30]. Finally, as monitoring is a measurable quantity, using models of trust, such as Taddeo's or Loi et al.'s [30, 40], the levels of JW-trust can be explicitly measured.

In other words, we want to direct attention to two features that explainability is supposed to have, in order to contribute to trust in the model sketched up this point. First, it is supposed to be an indicator of the reliability of the trusted entity, contributing to the justification of the user's belief that it is trustworthy. Second, it is supposed to justify a relaxation of monitoring of the trusted entity.

The model is fully general, but we start with the simplest possible account of the explainability-trust relation. In this account, the trusted entity is the AI *in itself*; its trustworthiness is understood as reliability. Thus, explainability could contribute to trust in AI by being a reliability indicator of the AI. In order to do so, it must contribute to the *justification* of the belief that the AI is trustworthy and, through that belief, lead to no or low monitoring levels. As we shall argue in Section 7, we do not believe that either condition is plausible. Representing the reliability of an algorithm is not what explainability *does*. And to the extent that it *contributes* to reliability, it can only do it through increased levels of monitoring.

7 EXPLAINABILITY FOSTERING TRUST IN A MEDICAL AI: IS THAT REALLY FEASIBLE?

In this section, we briefly comment on the pragmatic conditions under which explainability can actually support a physician's paradigmatic JW-trust in a medical AI. Our argument considers three points: 1) the different stages of interaction with a medical AI, 2) the spectrum of activities comprising monitoring a medical AI, and 3) the relation between holding a justified belief on the trustworthiness of the AI and the need of monitoring.

First, we argue that the interactions between a physician and a medical AI fall into two main categories, that we call *in-vitro* and *in-situ* for the sake of exposition. *In-vitro* interactions take place before the use of the medical AI in clinical practice. *In-vitro* activities consist of assessing the performance of the AI on test cases and comparing it with the one achieved by physicians [17], using the AI for training end-users or in empirical studies [26, 48]. *In-situ* refers to the class of activities that occur when the medical AI is fully deployed, i.e., in clinical practice. During *in-situ* interactions, a medical AI supports physicians in their decision-making by delivering outputs for their consideration in discussions with patients or other healthcare professionals.[22] We emphasize the distinction between *in-vitro* and *in-situ* as we consider the process of formation of (justified) beliefs on trustworthiness of a medical AI and its monitoring. In an ideal scenario, a physician undergoes multiple *in-vitro* interactions with a medical AI, possibly developing a mental model of selected AI functionalities, thanks to the use of explainability methods or reliability indicators, among others.[23]

[21] The reduction is with respect to the alternative scenario where no explanation of the AI "inner working" or outcome is provided. In this definition of warranted paradigmatic trust, the attitudes (i.e., refraining from monitoring) characterizing trust are required to be caused by, and not merely being coincidental with, beliefs about trustworthiness. A physician who happens to have a justified true belief in the trustworthiness of a medical AI, e.g., by reading about its test performance in an academic journal, but who is not free to choose whether to supervise it, control it, or test its outputs, and who wishes he could do it instead, does not have warranted paradigmatic trust in the AI, as we define it here. The physician merely happens not to closely monitor the AI, but for reasons that have nothing to do with his beliefs about its trustworthiness.

[22] In this paper, we consider two modes of interaction with medical AIs in *in-situ* relations: conventional and integrative [8, 45]. In conventional interactions, the AI collects patient's information from electronic health records (EHRs) as inputs (e.g., under the request of a physician) and returns an outcome. In integrative interactions, the AI can autonomously collect patient's information from the EHRs, share predictions with a physician, and update EHRs accordingly. In both modes of interaction, physicians make the final decision.

[23] If explainability (which leads to the understanding of errors) leads to debugging which improves the reliability of the AI when used *in-situ*, it may contribute to *high* warranted trust. [24] But this contribution is *indirect*. Explainability contributes to reliability in the debugging phase; its outcome is an *improved model*. From that point on, it is *only* the reliability of the improved model that justifies and warrants trust (and reduced monitoring). There is no more *direct* contribution of explainability to (user) trust. The (uncritical) *AI user* is not involved in any further recognition of errors. For example, the physician is not interested in explainability methods. He only cares that the AI is accurate, etc. Whether explainability methods have been used in debugging is entirely irrelevant to him. (Conversely, if the user's attitude *is* critical and leads to the direct engagement with explainability methods, our main argument applies.)

Moreover, we need to consider that monitoring a medical AI throughout an *in-situ* reliance relation may involve activities that aim at validating a given AI outcome,[24] but that do not fall within the scope of explainability methods. Examples of such monitoring activities include discussing an AI outcome with colleagues, patients, surrogates or other experts (e.g., data engineers), debugging the data-model-outcomes system to analyse normative aspects arising from the use of the medical AI, such as accountability [14], discussing possible biases in data used for training the AI, assessing data quality and performance trade-offs, and conducting an error analysis.[25] Therefore, considering all possible *in-situ* interactions of a physician with a medical AI, explainability is neither necessary nor sufficient for justifiably reducing the need to monitor the system.

Lastly, as shown in Table 1, a physician's belief on the trustworthiness of a medical AI and his need for monitoring are neither independent of each other nor *necessarily* related by a relation of justification. And as argued, the fact that AI is explainable does not logically entail that a reduction of monitoring activities is justified. Even if the reduction of monitoring occurs (our *a priori* model cannot make empirical predictions), it is not necessarily justified.[26] In other words, the gain in knowledge or understanding produced by explainability methods does not rationally warrants a reduction of the monitoring activities that ought to be in place in the AI-user interactions. On the contrary, one process in which explainability can be useful is the *critical* analysis of the outputs, e.g., deciding, for any suspicious AI output, whether it should be considered as a valid case or as possible error of the AI, to be tested by explainability methods. This practice contributes to the trustworthiness of the AI-as-used-by-a-human (to which we refer as "AI-user dyad"), but not to the reliability of the AI *in and by itself*, which remains, clearly, unaffected. (We turn to this distinct, but equally important, level of description of trust in Section 8.) And since the practice in question (validating outputs) is a form of monitoring *of* the AI *by* the user, it clearly does not contribute to the user's trust in the AI, if trust implies low or no monitoring.[27] This, of course, is not tantamount to predicting that explainability methods will not lessen the perceived need for monitoring an AI in some (misguided) physician. Empirically speaking, this may happen, and our *a priori* analysis cannot exclude it. But if the physician lowers monitoring levels, due to explanability, in a way that is not justified by those explanations, the resulting (paradigmatic) trust would not count as justified and warranted.

8 HOW EXPLAINABILITY FOSTERS TRUST IN AI: THE CASE OF AN AI-USER DYAD

We now provide an alternative perspective under which explainability supports trust in AI. Up to the present point, it has been assumed that the focus was trust of a human user in an AI, which is justified by the reliability of AI, considered as an agent-independent algorithm, correctly represented by a human user. But in most concrete circumstances, AIs do not take decisions alone. Rather, these systems are used to take decisions in an interactive process where a human being, the user, intervenes. The reliability of the resulting process is affected by the reliability of the AI, but also by the reliability of the process of use of the AI by the human. It is therefore possible to ask what gives one reason to trust the dyad "AI-user" as distinguished to trusting the AI in isolation. With respect to the dyad, it is possible to ask whether a given AI-user dyad is more reliable than another and because of which specific features of different users. Moreover, what would make the use of an AI reliable in the hands of this or that user, specifically? However, it is also possible to ask, whether AI-user dyads are, on average, more reliable when the AI is explainable to the user, compared to when it is not. The dyad idea assumes that the paradigm of use of an AI is not one of replacement of human intelligence, but one of augmentation. In the perspective of AI as "augmented intelligence," it makes sense to ask whether the public should trust it. To trust augmented intelligence is equivalent to trusting AI-user dyads.

If we assume that the AI-user dyad is the relevant unit of analysis, nothing prevents us from applying the interpretation of "explainability supports trust in AI" developed above to the idea that explainability supports trust in AI-user dyads. It may turn out that this interpretation provides a more plausible way to connect explainability with reliability and the justified belief in AI trustworthiness. We summarize this new interpretation in Table 2.

In summary, we have shown that there exist two explanations of the claim "explainability supports trust in AI," based on the same account of the conceptual relations between reliability, justified belief, and monitoring. Obviously, we are dealing with an ambiguous claim, because the claim can relate to the trust of the user in an AI, or the trust of the public in the AI-user dyad.[28] Simply put, the object of trust in the two cases differs and, in particular, the role of explainability vis-à-vis the object of trust differs, as well.

We argued that, when the object of trust is the AI, explainability can only contribute to trust by making the AI more reliable in such a way that it (justifiably) causes low monitoring levels. This seems the least promising interpretation to us, because established validation methods of reliability, which are epistemically independent of explainability, exist and are typically followed.

However, when the object of trust is the AI-user dyad, i.e., the user is not the trustor, but an element of the trustee, one can make sense of the idea that explainability contributes to trust. For clearly, the explainability of AI affects the relation between the user and

[24] Relying on the medical AI means to delegate the goal of generating a reliable prediction for a prognosis, diagnosis or treatment to the AI. The reliance relation successfully ends when the physicians accepts the AI prediction and makes use of it in shared decision-making. Therefore, monitoring encompasses all activities that aim at validating the predicted AI outcome. However, reliance is interrupted if the AI prediction is rejected.

[25] As noted before, a "Critical/careful/zealous physician" may invest in monitoring, even if he is entertaining a justified belief on the trustworthiness of the AI.

[26] This could be the case of a physician forming trustworthiness beliefs due to the causal influence of irrelevant factors such as the GUI appearance [24].

[27] It may be objected that the idea that trust means non-monitoring seems importantly time sensitive. Monitoring suggests lack of trust, but after a trustor monitored for a while the very act of monitoring can increase his trust, if monitoring suggests things have gone well so far. In reply, monitoring performed in past interactions cannot increase retrospectively the level of trust in those interactions. Explainability can increase trust in future interaction but only if it has led to a reliability gain during debugging. This gain must be assessed via computational evaluation and verification methods, that are deemed justified on *independent* epistemic grounds. Thus, the contribution of explainability to justification is indirect and inessential.

[28] The list is not exhaustive. The sentence could also be used to relate to the public trust in the companies producing the AI, and so on. Generally speaking there are too many possible trust relations for us to be able to map them all here.

Paradigm	**Artificial intelligence**	**Augmented intelligence**
Trustee (object of trust)	Algorithm.	AI-user dyad.
Trustor	User (e.g., physician).	Public (e.g., patient).
Reliability	The algorithm is reliable, e.g., accurate, precise, robust, etc. (as required by "contracts") *in and by itself.*	The algorithm as used by the human leads to outcomes that are accurate, precise, robust, etc. (as required by the relevant public and institutions).
Trustworthiness belief	The AI user (e.g., a doctor) believes that the AI output is reliable (that is, sufficiently accurate, precise, etc., given the relevant "contracts").	The public (e.g., patient) believes that the diagnosis provided by the AI-enhanced physician is reliable enough given public/institutional demands.
Justification of the trustworthiness belief	The belief (above) happens to be correct, because the AI output is generated by a reliable process (the algorithm is reliable, to the degree specified by "contracts").	The belief (above) happens to be correct, because the augmented intelligence is reliable (the process of a physician making a diagnosis with the assistance of AI is reliable, to the degree specified by the relevant public and institutions).
Causation	The belief (above) happens is caused by the reliability of the AI and causes a justified level of monitoring of the AI.	The belief (above) is caused by the reliability of the AI-user dyad and causes a justified level of monitoring of the AI-user dyad.
Monitoring attitudes	The user (e.g., physician) reasonably lowers the monitoring of the AI, because he believes that it is very reliable (precise, robust, accurate, etc). Given the objective reliability of the algorithm, this attitude is justified.	The public (e.g., patients) reasonably lower the monitoring of the AI-user dyad, which is justifiably believed to be reliable.

Table 2: This table compares the paradigm of artificial intelligence with that of augmented intelligence in relation to different trust-relevant elements. We maintain that the attitudes in the paradigm of augmented intelligence are more likely to occur and be justified *because* of explainability than those in the artificial intelligence one.

the AI. Even a reliable process can produce mistakes from time to time (this is entirely consistent with the AI being reliable to a high degree). Thus, it is to be expected that a physician may sometimes decide to doubt the output of the AI that he trusts moderately, but not completely. The attitude we describe here is the one of the "Critical/careful/zealous physician" in Table 1. A physician who understands the AI better (thanks to explainability methods) may be more likely to doubt the output of the AI when it is worth doubting, and to trust it when it is worth trusting. Since some AI-made errors may be corrected, explainability, which affects understanding, would then improve the reliability of the AI-user dyad.

Finally, let us consider the case of a physician, let us say a dermatologist, with established experience in diagnosing skin cancer cases. The physician consults an AI that classifies images of skin lesions in cancer vs. no cancer cases before making a diagnosis. Suppose that the AI and the physician arrive at opposite predictions for a medical image of a given patient. Then, the physician is tempted to discard the prediction of the AI, but he has reasonable doubts about this procedure due to the fact that the accuracy of the AI is, on average, higher than that of physicians with his degree of experience. Now, let us suppose that the physician consults an explainable AI method that highlights on the skin lesion image those sets of pixels that contributed the most to the prediction of the AI that is in question. In other words, the physician performs (much) monitoring of the case at hand. As a result, the physician notes that the important sets of pixels in question are far away from the skin lesion under examination. Therefore, he argues that it is highly implausible and unlikely that there exists a causal relationship between them and the nature of the skin lesion under observation. (Clearly, the physician appeals to his own background knowledge to arrive at this judgment.) Observing the output of the explanation reinforces his belief that in this case it is preferable to follow his own judgment and ignore that of the AI. Thus, if we understand trust as a relation between the patient and the AI-physician dyad, we can explain why an understanding of the source of error of AI contributes to the making the use of AI by a human more reliable in overall effect.[29]

Let us now reply to the following objection. Imagine that I am a medical patient, and suppose I learn that some AI, say, during its development, had been examined via explainability methods to test and calibrate its accuracy. In such a case, one would think, this would precisely be the kind of thing that would give me reason to trust the algorithm in a clinical setting. Why is such a story about how "explainability can foster trust" not admissible? Notice that this is not trust in the "AI-user dyad." Rather, it is trust in the AI making decisions by itself, but the trustworthiness stems from the fact that the algorithm was tested significantly before it was implemented, also thanks to explainability methods.

In reply, the use of explainability methods is not necessary for assessing the reliability of the AI, since reliability can be finally assessed independently of explainability. Even if explainability provides AI testers with useful heuristics that make debugging the

[29] Alvarado claims that understanding the nature, distribution, and source of error is key to the reliability of an AI and expresses skepticism with regard to the ability of physicians to detect errors in AIs [2]. In our interpretation, explainability could be key to the reliability of the AI-physician dyad if, and in so far as, it enhances the ability of physicians to detect AI errors.

AI easier, eventually explainability is not a necessary element for assessing reliability. As a matter of fact, many reliability indicators that are independent of explainability can be made available to both physicians, and, in principle, the public, after debugging. It may be psychologically harder for physicians to accept such indicators in the absence of explanations, but treating an explainable model tested on limited datasets as more reliable (and less in need of controls) than one that is non-explainable but proved successful in many more and more diverse cases is not justified. By contrast, the reliability of any given AI-physician dyad cannot at the moment be assessed by considering large amounts of test data – as such data is harder to produce and evaluate. And even if the relevant data will be produced over time, it is unlikely to be made available to patients, due to its sensitive nature. (An unreliable physician will attract social stigma for himself and the institution.) So explainability provides, in the dyad case, a meaningful proxy for reliability.

9 CONCLUSIONS

We have provided a philosophical explanation of the relation between AI explainability and trust in AI commonly referred to as "explainability fosters trust in AI." This explanation rests on considering justified beliefs on the trustworthiness of the AI and the relations of both justification and causality between such beliefs in the mind of the AI user and the monitoring activities during the use of AI. This is equivalent to requiring that explainability supports justified and warranted paradigmatic trust in AI. We argued that our approach of explainability and trust as anti-monitoring is able to intercept the complexity of the interactions between physicians and medical AI systems, providing a description of cases where humans hold different beliefs on the trustworthiness of the medical AI and exercise varying degrees of monitoring. We discussed a sketch of the empirical conditions under which explainability fosters *user's* trust in AI in clinical practice. We argued, based on *a priori* considerations, that the claim is implausible if the trust in question is paradigmatic JW-trust.

In an alternative description of the trusted entity, however, our account can explain why it is plausible that "explainability fosters trust in AI" and specify the empirical conditions for this to happen. It may foster trust in AI by fostering the trust of various different publics, i.e., third parties beside the AI and its user. In an augmented intelligence paradigm, what these various publics are asked to trust are not AIs, but AI-user dyads.

Thinking in terms of AI-user dyads can explain the apparent paradox that in order for the public to trust AI, the public must be able to trust (at least some) AI users not to trust it (fully). Explainability may contribute to *public* trust precisely thanks to this possibility, which, if levels of analysis are kept separated, is not a contradiction after all.

We argue that empirical research is needed to test the proposed meaning of "explainability fosters trust in AI." According to the account offered here, researching trust in AI in relation to explainability should focus on measuring 1) the perceived need for monitoring (both by user towards AI and by the public towards the AI-user dyad), and 2) the strength of the belief in trustworthiness (of both AI users and of the public).

If paradigmatic JW-trust is in question, then, trustworthiness must be investigated as well. In particular, the question whether AI explainability contributes to the trustworthiness of the AI-user dyad can be studied by measuring the reliability of the dyad, which may be operationalized as the accuracy, robustness, etc. of decisions resulting from the use of explainability methods by the AI users. Finally, new methods have to be designed to study, from an empirical point of view, if the trustworthiness of AI-user dyads and public trust in AI are indeed causally related. This is necessary to determine if public trust in AI is not only justified (due to valid proxies, such as brand reputation, etc.) but also warranted. If public trust is generally justified, but it can be not warranted, new auditing institutions for AI-user dyads, responding to different publics, may be designed. However, the question whether these control institutions are compatible with a high level of public trust in AI should be considered, as well.

ACKNOWLEDGMENTS

We thank all the anonymous reviewers whose comments helped improve this manuscript and its previous versions. The authors declare no additional sources of funding.

REFERENCES

[1] Ramón Alvarado. 2021. Explaining Epistemic Opacity. (2021). http://philsci-archive.pitt.edu/19384/

[2] Ramón Alvarado. 2022. Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI. *Bioethics* 36, 2 (2022). https://doi.org/10.1111/bioe.12959

[3] Annette Baier. 1986. Trust and antitrust. *Ethics* 96, 2 (1986), 231–260.

[4] Annette Baier. 2013. What is trust? In *Reading Onora O'Neill*. Routledge, New York, NY, USA, 185–195. https://doi.org/10.4324/9780203758793-21

[5] Nikola Biller-Andorno, Andrea Ferrario, Susanne Joebges, Tanja Krones, Federico Massini, Phyllis Barth, Georgios Arampatzis, and Michael Krauthammer. 2021. AI support for ethical decision-making around resuscitation: Proceed with care. *Journal of Medical Ethics* (2021).

[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[7] Jens Christian Bjerring and Jacob Busch. 2021. Artificial intelligence and patient-centered decision-making. *Philosophy & Technology* 34, 2 (2021), 349–371.

[8] Matthias Braun, Patrik Hummel, Susanne Beck, and Peter Dabrock. 2021. Primer on an ethics of AI-based decision support systems in the clinic. *Journal of Medical Ethics* 47, 12 (2021), e3–e3.

[9] Christiano Castelfranchi and Rino Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley, Hoboken, NJ, USA.

[10] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber. 2013. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. Springer, Berlin, Germany, 411–418. https://doi.org/10.1007/978-3-642-40763-5_51

[11] Partha Dasgupta. 1988. Trust as a commodity. Trust: Making and Breaking cooperative relations. D. Gambetta.

[12] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[13] Juan M Durán and Nico Formanek. 2018. Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines* 28, 4 (2018), 645–666.

[14] Juan Manuel Durán and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47, 5 (2021), 329–335.

[15] Catherine Z Elgin. 2005. *Non-foundationalist epistemology: Holism, coherence, and tenability*. na.

[16] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. 2021. What is interpretability? *Philosophy & Technology* 34, 4 (2021), 833–862.

[17] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (Feb 2017), 115–118. https://doi.org/10.1038/nature21056

[18] Andrea Ferrario, Michele Loi, and Eleonora Viganò. 2020. In AI we trust Incrementally: a Multi-layer model of trust to analyze Human-Artificial intelligence interactions. *Philosophy & Technology* 33, 3 (2020), 523–539.

[19] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262.

[20] Edmund L. Gettier. 1963. Is Justified True Belief Knowledge? *Analysis* 23, 6 (1963), 121–123. https://doi.org/10.2307/3326922 Publisher: [Analysis Committee, Oxford University Press].

[21] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316, 22 (2016), 2402–2410.

[22] Jianxing He, Sally L Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine* 25, 1 (2019), 30–36.

[23] Paul Humphreys. 2009. The philosophical novelty of computer simulation methods. *Synthese* 169, 3 (2009), 615–626.

[24] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[25] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[26] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[27] Arnon Keren. 2014. Trust and belief: A preemptive reasons account. *Synthese* 191, 12 (2014), 2593–2615.

[28] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* 46, 1 (Mar 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[29] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[30] Michele Loi, Andrea Ferrario, and Eleonora Viganò. 2020. How much do you trust me? A logico-mathematical analysis of the concept of the intensity of trust. *SSRN Electronic Journal* (2020).

[31] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.

[32] Carolyn McLeod. 2021. Trust. In *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[33] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[34] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[35] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.

[36] Onora O'Neill. 2002. *A Question of Trust: The BBC Reith Lectures 2002*. Cambridge University Press. Google-Books-ID: h_rTsfy4srQC.

[37] Alvin Plantinga et al. 1993. *Warrant: The current debate*. Oxford University Press, USA.

[38] Nadine Schlicker and Markus Langer. 2021. Towards Warranted Trust: A Model on the Relation Between Actual and Perceived System Trustworthiness. In *MuC '21: Mensch und Computer 2021*. Association for Computing Machinery, New York, NY, USA, 325–329. https://doi.org/10.1145/3473856.3474018

[39] Matthias Steup. 2004. Internalist reliabilism. *Philosophical Issues* 14 (2004), 403–425.

[40] Mariarosaria Taddeo. 2010. Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and machines* 20, 2 (2010), 243–257.

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

[42] Suresh Venkatasubramanian and Mark Alfano. 2020. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 284–293.

[43] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[44] David S Watson and Luciano Floridi. 2021. The explanation game: A formal framework for interpretable machine learning. In *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, 185–219.

[45] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[46] Linda Zagzebski. 1994. The inescapability of Gettier problems. *The Philosophical Quarterly (1950-)* 44, 174 (1994), 65–73.

[47] Carlos Zednik. 2021. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology* 34, 2 (2021), 265–288.

[48] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852