Ethics and Information Technology (2025) 27:60
https://doi.org/10.1007/s10676-025-09867-w

ORIGINAL PAPER

Trust and Artificial Intelligence in the Doctor-Patient Relationship: Epistemological Preconditions and Reliability Gaps

Eike Buhr[1] · Orhan Onder[2,3] · Pranab Rudra[4] · Frank Ursin[4]

Abstract

While trust is foundational to the doctor-patient relationship, the introduction of AI into healthcare settings poses the risk of eroding this trust, and such erosion cannot be countered simply by appealing to the notion of "trustworthy AI." We argue that trust presupposes specific epistemic attitudes that cannot be meaningfully applied to AI systems. Accordingly, our focus is not on specifying which capabilities AI must exhibit in order to appear trustworthy, but on examining from an epistemological perspective how the use of AI reshapes the dynamics of trust within the doctor-patient relationship. To this end, we first sketch conceptions of trust and demonstrate how trust differs from reliance. We then combine the model of Computational Reliabilism with an epistemic framework to develop a matrix for the ethical analysis of our use cases. Finally, we apply this framework to three scenarios of melanoma detection, risk prediction, and psychotherapy chatbots, which we construct by mapping epistemic stances across different modes of human-machine interaction, ranging from collaborative support with varying degrees of autonomy to the replacement of human-human interaction. We argue that the application of AI in the doctor-patient relationship exposes what we call a "reliability gap" — a conceptual space where the opaque nature of advanced AI systems prevents both doctors and patients from independently verifying their reliability. This creates a dynamic where reliability in the AI's performance is increasingly mediated by the doctor as a proxy. Our use cases demonstrate that the more autonomous and opaque AI systems are, the more trust in the doctor becomes essential for bridging reliability gaps, while threatening to overburden the doctor's central role.

Keywords Artificial intelligence · Trust · Healthcare · Doctor-patient relationship · Ethics · Epistemology

Introduction

AI-based technology is increasingly being used in healthcare where it has already demonstrated significant success in terms of its performance (Topol, 2019). However, its use also carries the risk of undermining essential relational dynamics, particularly the trust between doctor and patient. While doctors establish trust by e.g., being knowledgeable, displaying skill, communicating openly and showing empathy (Wu et al., 2022), these aspects of a fiduciary relationship are challenged by AI's functioning. For example, opaque algorithms, lack of explicability and human oversight, privacy concerns, de-skilling of healthcare professionals, and transparency about whether AI is being applied at all complicate the establishment of trust between doctor and patient (Davenport & Kalakota, 2019; Grote & Berens, 2019; Sauerbrei et al., 2023). To establish the conditions for trust also in the AI-supported doctor's office, the EU's High-level Expert Group on Artificial

✉ Eike Buhr
eike.buhr@uni-oldenburg.de

Orhan Onder
orhnonder@gmail.com

Pranab Rudra
pranab.rudra@stud.mh-hannover.de

Frank Ursin
ursin.frank@mh-hannover.de

[1] Ethics in Medicine, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

[2] History of Medicine and Ethics, Marmara University School of Medicine, Istanbul, Istanbul, Turkey

[3] Philosophy of Media and Technology, Department of Philosophy, University of Vienna, Vienna, Vienna, Austria

[4] Institute for Ethics, History and Philosophy of Medicine, Hannover Medical School (MHH), Hannover, Hanover, Germany

Published online: 22 October 2025

Intelligence (HLEGoAI) has formulated criteria to ensure the trustworthiness of AI systems (HLEGoAI, 2019). It is, however, questionable whether the abovementioned epistemic challenges can be resolved by referring to trustworthy AI (London, 2019). While trusting AI primarily relates to predictive reliance in terms of assuming technical functioning and thus the adoption of a certain epistemic position towards the way information is processed by AI, trusting a doctor often involves affective and emotional attitudes such as empathy (Jones, 1996), though other accounts emphasize normative expectations or role-based commitments rather than affective motives. Since the fiduciary doctor–patient relationship presupposes a notion of trust grounded in the assessment of attitudes and intentions, whereas extending trust to AI-based technology instead engages distinct cognitive operations (Malle & Ullman, 2021), it remains an open question whether AI systems can be trusted in the same way as human agents.

We hypothesize that the use of AI might eventually compromise the trust relationship between patient and doctor and cannot be mitigated by pointing to trustworthy AI. Recent empirical work offers support for our hypothesis: In an online vignette study, participants trust a human doctor more than a hybrid team of a human doctor with an AI system or even an AI system alone (Riedl et al., 2024). A qualitative study found that patients have certain information needs for trusting relations such as how AI tools are overseen, how they impact their care, and how physicians use them (Stroud et al., 2025). In a mixed methods study by (Shevtsova et al., 2024), epistemic factors such as clinicians' knowledge about AI, performance and explainability of AI were identified as among the most relevant technology-related factors for trust among stakeholders in medicine. Particularly for rare, complex and high-risk cases (Zondag et al., 2024), clinicians have difficulties trusting "black box" models that lack interpretability (Nouis et al., 2025).

We aim to show that trust requires specific epistemic attitudes that cannot be meaningfully applied to AI systems. Although the use of AI in clinical practice may appear trustworthy through its embedding in specific socio-technical contexts, this perceived trustworthiness does not extend to the AI system itself. The opacity of many AI models generates "reliability gaps," insofar as their performance cannot be adequately assessed or understood by either patients or physicians. While such gaps can, to some extent, be compensated for by placing trust in the treating physician, this dynamic ultimately risks overburdening the doctor–patient relationship.

A critical epistemic limitation of AI in clinical contexts is algorithmic opacity, which refers to the lack of transparency in how AI systems reach their conclusions. Complex machine learning models, such as deep neural networks, often operate as "black boxes," making it difficult for clinicians and patients to understand or explain the reasoning behind specific outputs (Burrell, 2016). This opacity can lead to challenges for accountability, in case it is unclear how to justify an AI-based decision that is potentially erroneous, imposes a bias or does harm to people (Pasquale, 2015). The HLEGoAI states that "[e]xplicability is crucial for building and maintaining users' trust in AI systems" (HLEGoAI, 2019, p. 13). To counter the epistemic opacity as an epistemic limitation of current deep learning (DL)-based AI systems appears to necessitate explicability to establish trust. Ironically, if a system is fully explicable, there is no need for trust; trust becomes necessary only when a system is not explicable, as it is required in situations of heightened epistemic vulnerability.

Against this backdrop, we argue that while trust between doctor and patient is grounded internally, i.e., by referring to the epistemic and professional capacities of the doctor as perceived by the patient, trust in the use of AI is grounded externally, i.e. by referring to the socio-technical context and developers of Jacovi et al. (2021) capture this perceived distinction by differentiating between intrinsic trust, which arises when users believe they understand and align with AI's reasoning, and extrinsic trust, which arises from external assurances such as evaluation data or institutional safeguards. While we maintain that AI itself lacks the capacities required to be a genuine object of trust, their framework demonstrates how reliance on AI systems is socially and institutionally mediated. If the trustworthiness of AI-usage relies on its operation within a trustworthy socio-technical environment (HLEGoAI, 2019), the question inevitably arises as to what extent the healthcare system provides such an environment. As doctors and other healthcare workers are usually no experts in AI-systems, we ask to what extent AI-usage in the healthcare sector is perceived as trustworthy and in which way the implementation of AI influences the fiduciary doctor-patient relationship in different use-cases.

To address this question, some preliminary conceptual clarification is required. Therefore, the following section examines the conceptual foundations of trust and explores how it diverges from the notion of reliability. In doing so, we identify critical conditions within the socio-technical environment that enable the use AI systems to be perceived as trustworthy. The discussion culminates in the recognition that the trustworthiness of AI-usage is context-dependent, shaped by factors that vary significantly across different applications. For example, the trust underpinning the interaction between a doctor and a patient using AI for melanoma detection is based on different criteria compared to a psychotherapy chatbot, where the doctor's role is largely absent. In the subsequent section, we analyze these contextual factors through three distinct use cases. Here, we consider an AI system for distinguishing between melanoma and nevi, a predictive system for Alzheimer's dementia, and a therapeutic chatbot. These examples represent diverse modes of human-machine interaction, offering a framework for examining which factors are necessary for the use of AI to be perceived as trustworthy and how the use of AI influences trust within the doctor-patient relationship. The key observation here is that AI use appears trustworthy when its reliability can be dependably assessed. For assessing reliability, we use the model of Computational Reliabilism for AI in medicine proposed by Durán and Formanek (2018) and Durán and Jongsma (2021).

The analysis unfolds in two stages. First, we investigate the epistemic processes patients must engage in to regard the use of AI as trustworthy and the normative basis on which epistemic authority can be appropriately attributed (Keren, 2007). Here, we examine the epistemic foundations of trust, contrasting the cognitive operations underlying the assessment of technical functioning of AI-based technologies with the interpersonal dynamics inherent in the traditional doctor-patient relationship to show how each situation might cause a different kind of epistemic asymmetry (Onder, 2022). Our analysis suggests that as AI systems become increasingly opaque and autonomous, the evaluation of their reliability progressively merges with trust placed in the doctor. This convergence also raises important ethical considerations. Here, we identify scenarios where the reliability of AI systems cannot be assessed by doctors nor patients, thus leading to *reliability* gaps, ultimately complicating trust within the doctor-patient relationship. Second, we explore whether the use of AI undermines or, under certain conditions, enhances the trust-based relationship between doctor and patient.

To better understand the concept of trust regarding the application of AI in the context of doctor-patient relationship, we propose a matrix to encompass these diverse perspectives and illustrate how the assessment of AI-use in healthcare regarding its influence on trust encompasses various epistemic dimensions, ranging from source of knowledge to reasoning and decision-making (s. Table 1). By mapping these dimensions, we elucidate the complex interplay between AI's technical capabilities and epistemological dimensions, providing a context-sensitive lens for examining ethical aspects of trust in AI-mediated healthcare. We conclude with a broad concept of trust that links different conceptions of trust and their epistemological preconditions and systematically demonstrates their ethical implications. In this regard, we demonstrate that the use of AI can appear trustworthy but not the AI-system itself. Doing so, we aim to show why it is morally relevant to consider the epistemic foundations of trust when discussing medical AI. Thus, this article not only illustrates the interconnectedness of epistemology and bioethics but also provides a starting point for future empirical and psychological work on the question.

Philosophical Foundations of Trust and Reliance

Since the term trust is used quite broadly in everyday language, defining this key concept is mandatory. To begin with, trust is usually directed to a specific domain (McLeod, 2023), for example, we trust our general practitioner to take good care of our medical needs. Accordingly, "trust is generally a three-part relation: *A* trusts *B* to do *X*" (Hardin, 2002, p. 9). Further, trust is often considered an interpersonal phenomenon, but we can also trust groups of people, e.g., doctors in general, or organizations, e.g., the healthcare system (Hall et al., 2001).

Although the precise meaning of trust is philosophically contested, most theories stress "the *optimistic* acceptance of a *vulnerable* situation in which the truster believes the trustee will *care* for the truster's interests" (Hall et al., 2001, p. 615): "trust is inseparable from vulnerability, in that there is no need for trust in the absence of vulnerability". McLeod (2023) states that „[t]rusting requires that we can, (1) be vulnerable to others – vulnerable to betrayal in particular; (2)

Table 1 A matrix for evaluating reliability of AI systems

Categories	Source of know-ledge	Representation of knowledge	Management of knowledge	Inference methods	Reasoning and decision-making
Reliability and Validatability	Is the source of knowledge appropriate for the task?	Is the knowledge represented accurately?	Can the AI be applied to data beyond its training set?	How does the AI arrive at its results?	Has the AI drawn correct conclusions?
Robustness	Are the knowledge sources free from historical biases that may affect robustness?	Is the representation of knowledge sufficiently generalizable to support robust performance across varying contexts and datasets?	Has the AI been thoroughly tested? Can the AI be applied to data beyond its training set?	Does the AI produce consistent inferences and recommendations across varying conditions and datasets?	Does the AI deliver consistent decisions under varying conditions?
Imple-mentation history	Have similar knowledge sources in past applications supported consistent and reliable system performance?	Has the representation of knowledge been consistent and free from persistent structural biases?	Has the AI been consistently updated?	Has the AI's inferential process produced reliable and valid decisions in past applications?	Has the AI consistently produced reliable results in the past?

rely on others to be competent to do what we wish to trust them to do; and (3) rely on them to be willing to do it" (p. 4). This shows that trust goes beyond reliance.[1] Furthermore, it shows that someone's willingness and competence are the conditions that make a person trustworthy for a specific regard. Different understandings exist of what it means for someone to be willing to perform an action. In this regard, a broad distinction can be made between motive based and non-motive based theories of trust.

For motive-based theories of trust, it is not sufficient that the trustee is motivated to act in a certain way, the motivation must also be of a certain nature. Several proposals exist as to the nature of motivation that is relevant to trustworthiness. For example, Hardin (2002) suggested to locate the nature of the trustee's motivation in their self-interest, specifically in maintaining their relationship with the truster thus "encapsulating" the truster's interests within their own. In contrast, others have suggested that the relevant nature of the trustee's motivation consists in goodwill (Baier, 1986, 1991; Jones, 1999) which, roughly, means that the trusted person will act with a benevolent attitude or genuine concern for the truster's well-being. Moreover, proposals exist that locate the nature of the trustee's motivation in a moral commitment or obligation or in moral virtue (McLeod, 2023).

For non-motives-based theories of trust, trust is not grounded in the trustee's motivations but rather in the truster's normative expectations—what they believe they should be able to count on from the trustee. These theories aim to show that trust involves a distinct stance or belief in what is owed in a relationship, rather than a simple expectation of behavior, thus distinguishing trust from mere reliance (McLeod, 2023). For example, Holton (1994) suggested taking a "participant stance" by "treating the trustee as a person – someone who is responsible for their actions" (McLeod, 2023, p. 15). Building on this, Walker (2006) proposed a "normative-expectation" theory according to which trust and reliance can be differentiated by highlighting that trust involves normative expectations whereas reliance involves predictive expectations. For example, one trusts a friend to return a borrowed car because people are normatively expected to return what they borrow, whereas one relies on their car to start in the morning not due to any normative obligation but because it has the technical capacity to do so and has started successfully the previous morning.

A further prominent proposal of a non-motive-based theory of trust is the "trust-responsive" theory, endorsed by philosophers like Faulkner (2011, 2017) and Jones (2012, 2019), which holds that trustworthiness involves responding appropriately to the expectation that someone will act because they are counted on. Finally, according to the "commitment account" of trust by Hawley (2014, 2017) someone is willing to perform an action if they have a commitment to do so. Hawley (2014, p. 11) explains that "commitments can be implicit or explicit, […] conferred by roles and external circumstances [… and] that mutual expectation and convention give rise to commitment".

Consequently, having outlined different conceptions of trust, we find that despite their differences, trust encompasses reliance plus an additional factor (Durán & Pozzi, 2025; Hawley, 2014, p. 5). A key factor distinguishing trust from reliance is delegation: whereas reliance involves actual delegation, trust does not require it but is sustained by the possibility that the truster could delegate to the trustee if necessary, so the relationship endures even without such delegation (Blanco, 2025b). While reliance also involves expecting someone or something to (re-)act in a specific manner, it does not require affective or normative stance towards the truster or willingness on part of the trustee. As extensively discussed by many scholars, these conditions of trustworthiness are scarcely applicable to AI (Bryson, 2018; Hatherley, 2020). This prompts an examination of the underlying framework within which the trustworthiness of AI is debated (Reinhardt, 2023). A useful perspective on this issue can be derived from Gunkel's ontology, which explores the status of AI as a potential moral agent. Gunkel (2023) refers to the difference between "things" and "persons" in western philosophical and legal thought. As AI-systems are increasingly able to mimic or surpass human intelligence, act autonomously, and even simulating empathy in a convincing manner as well as most importantly act with moral consequences, Gunkel argues, that the line between "thing" and "person" becomes increasingly blurred. Against this backdrop, some researchers argue that albeit AI does not have the above-mentioned competencies, it still can appear trustworthy, as advanced AI systems are capable of language use and appear autonomous to some degree. For example, Blanco (2025b) expands the concept of trust to AI by redefining motives as system criteria, agency as quasi-agency, and rational reasons as performance- and process-based justifications, thereby extending trust beyond interpersonal contexts. Safdari (2025), in turn, grounds the possibility of trust in empathic relations, suggesting that humans perceive AI as "otheroids" and that trust arises from experiential openness and a history of interaction. Similarly, Coeckelbergh (2012) advocates for adopting a stance of "quasi-trust" towards AI and Taddeo (2011) argues for the concept of "e-trust". Such a form of trust is founded on success rates of the trustee in similar actions, e.g. with a threshold value

[1] One exemption are risk-assessment views of trust which do not distinguish between trust and reliance. For risk-assessment views, a trustee‘s trustworthiness depends on the risks associated with relying on them, which, for example, may depend on the probability of the trustee‘s motivation to act a certain way will endure (McLeod, 2023).

used to determine whether an agent can be considered trustworthy (Taddeo, 2010a, 2011). In comparison, Ryan (2020) calls such a concept of trust *rational trust*, or reliance. The last aspect further stresses that trust in AI is different from trust in a human being. The HLEGoAI regards transparency as essential for trustworthy AI, but explainability only as one possible means, neither necessary nor sufficient. Amann et al. (2022) adopt a similar view, while Ferrario and Loi (2022) further stress that explainability promotes trust only when it enables justified paradigmatic trust by reducing the need for monitoring.

From the perspective of the HLEGoAI, for AI to be perceived as trustworthy, several elements must align: not only the AI system itself, but also the developers, organizations, users, and the broader socio-technical framework in which AI is developed and deployed must demonstrate trustworthiness (HLEGoAI, 2019). To address this, the HLEGoAI outlines three essential components of trustworthy AI. First, AI must comply with existing laws relevant to its development, deployment, and use. These include regulations such as data protection laws and product safety standards. Second, AI must be both technically and socially robust, meaning that appropriate safeguards are in place to ensure that AI systems do not behave unpredictably. Third, AI must adhere to ethical principles. Here, the HLEGoAI identifies four key ethical principles rooted in fundamental EU rights, including dignity, freedom, justice, and equality: respect for autonomy, non-harm, fairness, and explicability. For instance, AI should not exert coercion and must allow for human autonomy in decision-making, with human oversight being a core element. In discussing fairness, the HLEGoAI acknowledges the complexity of the concept but highlights both substantive and procedural fairness. The goal is to prevent bias and stigmatization, ensure equal opportunities, and provide affected individuals with a means to contest AI decisions. Additionally, accountability must be clearly established to prevent gaps in responsibility ascription (Lang et al., 2023), and AI decisions should be explainable. However, the principle of explicability does not imply that AI systems can never operate opaquely, as this fourth ethical principle allows for certain technical complexities in AI functioning (Ursin et al., 2023).

While the principles developed by the HLEGoAI guide AI development and governance, they do not allow for answering the question of whether specific AI systems possess the competencies necessary to be deemed trustworthy in terms sketched above. Instead, they serve as a framework for shaping the responsible development and deployment of AI. These guidelines impose requirements on the technical robustness of AI systems and the responsible behavior of involved actors—developers, users, and institutions alike. Hence, we concur with Ryan (2020) when he describes the notion of establishing trusting relationships with AI as a "radical claim" that cannot be conceptually achieved (p. 2). While one can reasonably rely on AI to perform its functions based on its design and capabilities, it is inappropriate to extend trust in the same way we do towards humans. Unlike human trust, which endures even without actual delegation because it is grounded in the possibility of delegation, reliance on AI does not persist if delegation does not take place (Blanco, 2025b). Trust in humans presupposes moral agency, reciprocity, and accountability, qualities that AI lacks. Hence, trust in people and AI is based on fundamentally different epistemic attitudes. This means, the key shift in focus is no longer on whether we should trust AI, but on assessing the degree to which AI proves itself to be reliable. All approaches that discuss a potential trustworthiness of AI either argue for a quasi-trust or e-trust (Coeckelbergh, 2012; Ryan, 2020; Taddeo, 2010b, 2011), expand the notion of motives to include technical functioning (Blanco, 2025a), or refer to external factors, e.g., the AI's functioning, that are designed to contribute to the trustworthiness of AI (Durán & Jongsma, 2021; HLEGoAI, 2019). However, the reference to the functioning of AI only explains why AI is reliable. The reference to external factors, e.g., the socio-technical context, may explain why one might aptly trust the use of an AI-system in a given context by a given actor. Still, it does not explain why the respective AI-system itself is trustworthy. In the following, we therefore examine the conditions under which the use of AI in clinical practice may appear trustworthy, the requirements that AI systems would have to meet in order to sustain such an appearance, and how their use relates to the trusting relationship between doctor and patient in different scenarios. This analysis thus addresses not only the reliability of AI systems and the trustworthiness of their deployment, but also the implications of AI use for trust in the doctor–patient relationship.

Epistemic Foundations of Trust and Reliance: What Is Necessary to Know Whether Someone Is Trustworthy?

The use of AI has broader implications for the trust-based doctor-patient relationship, as it alters the dynamics of trust. The issue extends beyond the reliability of AI systems to encompass the trustworthiness of AI-usage and thus the responsible institutions and respective contextual factors. This shift demands different epistemic attitudes than those required for assessing the trustworthiness of individual doctors within the traditional doctor-patient framework. Therefore, in the following section, we will examine whether and to what extent these differing epistemic attitudes conflict with one another and what ethical consequences this may have for the doctor-patient relationship and the quality of patient care. When examining the epistemic foundations of trust, we argue that trust relations develop gradually by assessing if the trustee fulfills the above-mentioned conditions from the perspective of the truster. While these epistemic operations can be contextual and subjective, there are notable differences that establish the truster's epistemic state towards the trustee.

Several studies have argued that the doctor-patient-AI relationship should be understood as a triadic structure rather than a mere extension of the doctor-patient dyad (Lorenzini et al., 2023; Onder, 2025). The asymmetric nature of the doctor-patient relationship persists within this triad, shaping interactions among all three actors (Onder, 2025). One of the key aspects of this asymmetry is epistemic asymmetry, which plays a crucial role in the establishment of trust. The extent of this asymmetry — whether between the doctor and AI, the patient and AI, or across the entire triad — directly influences how trust relationships are formed and maintained. As the depth of epistemic asymmetry varies, it affects how knowledge is distributed, interpreted, and relied upon, ultimately shaping the dynamics of reliability, trustworthiness and trust in the triadic relationship. Since the depth of epistemic asymmetry is shaped by the epistemological differences among the actors, its evaluation must consider how these differences manifest across various medical contexts (Onder, 2025). Therefore, examining these differences is essential for fully understanding the nature of epistemic asymmetry in such environments. Evaluating the epistemic status and addressing differences can be guided by a framework that distinguishes between human and AI trustees (Onder, 2022).[2] Through this framework, epistemological differences between AI and doctors can be systematically analyzed using a matrix that compares different clinical scenarios across four key dimensions: source of knowledge, knowledge representation, knowledge management, inferential reasoning and decision-making.

a) **Sources of Knowledge**: Doctors acquire knowledge through formal education, clinical experience, and selective access to medical literature, integrating theoretical and practical expertise with contextual considerations from patient interactions (Montgomery, 2005; Patel et al., 2012). In contrast, AI systems rely on vast datasets analyzed through machine learning algorithms, offering rapid data processing but lacking the intuitive and context-sensitive judgment of human clinicians (Asan et al., 2020; Esteva et al., 2017).

b) **Knowledge Representation**: Doctors synthesize explicit and tacit knowledge, using holistic reasoning to integrate diverse information sources and navigate contextual nuances (Polanyi, 1966). AI systems represent knowledge as structured data within probabilistic models, excelling in correlation identification but struggling with ambiguity and inferential reasoning beyond their training data (Dreyfus, 1992).

c) **Knowledge Management**: Doctors continuously refine their knowledge through practice and ongoing education, enabling dynamic adaptation to new information and situations (Greenhalgh, 2009). AI systems require explicit updates and retraining to incorporate new data, which can introduce challenges like biases or outdated algorithms if not managed properly (Obermeyer et al., 2019).

d) **Inferential Reasoning and Decision-Making**: Doctors employ deductive, inductive, and abductive reasoning to form diagnoses and treatment plans, considering not only the logical outcomes but also the ethical implications of their decisions. This approach allows them to engage with patients on a personal level, adapting their communication style and decision-making process based on the patient's unique needs and preferences (Montgomery, 2005). AI systems follow rule-based, statistical reasoning, offering data-driven recommendations but lacking the flexibility, ethical judgment, and interpersonal engagement of human decision-making (Russell & Norvig, 2016; Topol, 2019). While doctors are able to communicate their reasoning and decision-making according to the patient's understanding and preferences, explain potential alternatives and thus, facilitate shared-decision making, AI-systems cannot be argued with as they simply present their result as a product of their calculations. It is especially this aspect, that marks the difference between doctors and medical AI-systems in terms of reasoning and decision-making and that serves as a criterion to distinguish between trust and reliance.

As argued above, AI systems themselves cannot be regarded as trustworthy, although their use may be. On the basis of the outlined framework, such use can be deemed trustworthy only if the reliability of AI functionality is dependably assessable across all dimensions. Based on these epistemological aspects, we can formulate criteria to assess whether an AI system is reliable, as this reliability constitutes a condition for its use being perceived as trustworthy, and how this may influence the fiduciary doctor-patient relationship. We refer to the model of Computational Reliabilism for AI in medicine proposed by Durán and Formanek (2018) and Durán and Jongsma (2021). According to the authors,

[2] Strictly speaking, since AI systems cannot themselves be trustworthy, one would need to refer to them as a *reliancee*. For the sake of simplicity, we refrain from introducing this neologism here.

for an AI system to be reliable, its results must be verifiable and validatable. This entails assessing whether the AIs model was correctly implemented and whether its results align with real-world data. This criterion encompasses all the epistemological aspects mentioned above, as it involves verifying whether the AI used appropriate sources of knowledge, represented them accurately, applied them to new data effectively, and drew correct conclusions. Furthermore, for an AI model to be reliable, it must exhibit robustness. This criterion can also be applied across all epistemological dimensions but mostly pertains to reasoning and decision-making to knowledge management, ensuring that the AI has been thoroughly tested and can be applied to data beyond its training set, such as evaluating whether the model is susceptible to overfitting. The third criterion relates to the implementation history of the AI model, which involves determining whether the system has consistently produced reliable results in the past. This criterion also reflects every epistemological aspect outlined earlier, as a positive assessment of implementation history is generally the outcome of proper knowledge management. Durán and Jongsma (2021) also identify expert knowledge as a criterion for AI reliability. However, we argue that this criterion does not concern the functionality of the AI itself but rather the experts using the AI. As such, it falls outside the scope of a framework for testing the reliability of a given AI system as a condition for its use being perceived as trustworthy. Instead, it pertains to the socio-technical context of AI development and application, addressing the trustworthiness of the human actors involved in this process. This is a further indicator that the employment of AI in healthcare threatens to overburden the actors involved as they usually are no experts in AI. As we aim to focus on the reliability itself and not on the socio-technical context, in the following, we focus solely on the three criteria of verification and validation (1), robustness (2), and implementation history (3). With these three criteria in mind, we now turn to three use cases to assess what epistemological conditions must be fulfilled, so that a patient or a doctor can assess the reliability of a given AI-system and its influence on the trust between doctor and patient. For a comprehensive overview regarding the framework for evaluating reliability of AI systems, see the table below (Table 1).

Three Use Cases of AI in the Doctor-Patient-Relationship

The question of how the reliability of an AI system can be assessed by a patient, or a clinician is inextricably linked to the type of human-AI-interaction that is imposed by the socio-technical design and function of a specific AI system. Hence, in this section we aim to map the different epistemological stances and operations that are present in the different forms of interactions between doctor, patient, and AI in relation to trust. There are several models to describe types of human-AI-interaction. These types depend on the AI systems' capacities compared to humans (narrow/weak AI outperforms humans on specific cognitive tasks and general/strong AI has human-level intelligence), the cognitive tasks they simulate (perception, reasoning, knowledge, planning, communication) or the type of analytics they perform (descriptive, diagnostic, predictive, prescriptive) (Mökander et al., 2023). Vernon (2014) draws a broad differentiation for artificial cognitive systems, distinguishing between cooperation (agents work parallel or serial without a common goal) and collaboration (agents share a common goal in problem-solving while interacting). In the context of collaboration, the framework of human-AI interaction proposed by Simmler and Frischknecht (2021) is grounded in the levels of automation (ranging from decision support to fully automated execution) and the degree of epistemic autonomy that an AI system can and should be afforded. They differentiate levels of autonomy according to dimensions of determination, transparency, adaptability and explainability.

According to this distinction regarding different kinds of human-machine interaction, we discuss three use cases, ranging from collaboration with different degrees of autonomy to replacement. Here, we analyse the epistemic foundations of trustworthiness in doctor-patient-AI relationships regarding the different epistemological aspects named above. A comprehensive overview is provided at the end of this section (Table 2). Based on this, we discuss ethical implications for the different use cases.

The first use case involves an AI-based system for melanoma detection. AI applications are already widely utilized in dermatology, with the performance of advanced systems often matching that of expert dermatologists (Willem et al., 2022). The specific system under discussion conducts dermatoscopy, identifying melanoma based on structural and border analysis (Winkler et al., 2021). It highlights various skin areas by their associated risk levels. According to the taxonomy by Simmler and Frischknecht (2021), this system is characterized as determined, transparent, and explainable. Consequently, the human-machine interaction is collaborative: the doctor uses the AI for a specific task, maintains oversight of its operation, and retains full decision-making authority. In terms of reliability, the system's outputs are verifiable and validatable. Doctors can confirm the AI's findings and understand the criteria it facilitates to distinguish melanoma from benign naevi. This transparency allows doctors to assess the reliability of the system's knowledge sources and representations. Additionally, doctors can evaluate the system's robustness by applying it to

Table 2 Evaluation of Reliability of Three AI Systems and Their Influence On Trust Towards the Doctor

Criteria AI systems	Reliability and validatability	Robustness	Implementation history	Trust towards doctor
Case 1: melanoma detection	√	√	√	Potentially enhancing trust (if reliable performance is verified)
Case 2: risk prediction	To a limited extent as accuracy of prediction can only be verified retrospectively	√	√	Influence on trust dependent on AI's performance (retrospective validation needed)
Case 3: chatbot	X	X	Plausibility of chatbot answers can be assessed	Good AI performance enhances trust while bad performance lowers trust (based on plausibility of answers only)

new datasets when treating patients. Such usage enables them to detect potential overfitting or biases in the dataset, as well as to identify when updates to the AI system are necessary. Moreover, the system's implementation history is accessible, enabling doctors to review its past performance and account for its reliability over time.

In this use case, all three reliability criteria as outlined in the framework above —verification and validation, robustness, and implementation history —can be effectively assessed by the doctor. The doctor can transparently communicate the AI's purpose, functionality, and results, thereby preserving their competence in the eyes of the patient. Criterion (1), the ability to verify the AI's results and validate its processes, is particularly critical for maintaining this perception of competence. Furthermore, the use of such a system can be interpreted as an expression of the doctor's commitment to meeting patient expectations and acting with benevolence. Thus, in this use case, the AI's reliability is rooted in its functional transparency and its accuracy, which in turn can be verified by the doctor. This reliability can be effectively communicated to the patient, fostering an understanding of the AI's purpose and operation. As a result, employing this AI system has the potential to enhance the trust-based relationship between doctor and patient as the doctor appears competent and willing to use advanced technology to improve the care of their patients.

The second use case pertains to a system designed to support the diagnosis of Alzheimer's dementia. Similar systems are currently being developed (Dyrba et al., 2018, 2021). Unlike most other psychiatric disorders, Alzheimer's dementia is characterized by a relatively well-understood pathomechanism, enabling early diagnosis based on biomarkers and imaging, even in asymptomatic individuals (Sperling et al., 2011; Ursin et al. 2021). The system in this use-case utilizes brain scan evaluations to determine the presence of Alzheimer's dementia, and it provides a percentage-based probability assessment. At this level of functioning, its influence on the trusting doctor-patient relationship is comparable to that of the melanoma detection system. To adapt the use case, we propose extending the system's functionality to include the capability of predicting the progression of Alzheimer's dementia over time. The system processes vast amounts of data and independently derives the criteria for predicting disease progression, rather than relying on predefined rules or doctor input. This enhancement renders the system more autonomous compared to the melanoma detection system. While it remains a single-task system, its operations become less transparent and less explainable.

Although the system continues to function in collaboration with the doctor, who retains final decision-making authority, the doctor's ability to fully oversee and understand the system's internal processes is diminished. In terms of reliability, the results generated by the system remain verifiable and validatable. However, the predictive nature of this functionality introduces a temporal challenge, as it pertains to future disease trajectories and thus, its accuracy cannot be immediately confirmed. The system's results can, therefore, only be verified to a limited extent. The method of knowledge representation and sourcing differs in that it not only provides a current risk assessment but also includes a prediction derived from probability calculations. While a doctor can always assess the plausibility of the information, the accuracy of the prediction can only be verified retrospectively. Due to the system's opaque functioning, the process by which the AI generates its predictions about disease progression can only be partially validated. Consequently, criterion (1) is fulfilled only to a limited degree. Criteria (2) and (3), however, are generally satisfied and can be evaluated, as the doctor is able to assess the system's applicability to new datasets and identify potential issues such as overfitting or bias. Nevertheless, the doctor cannot explain why the AI remains applicable to novel types of data. This limitation reduces the doctor's ability to fully ascertain the reliability of the AI system in predicting the progression of Alzheimer's dementia, necessitating trusting the socio-technical framework within which the AI is developed and deployed. This situation raises critical concerns. Should a doctor, bearing significant responsibility for decisions with profound implications, be expected to place a "leap of faith" in the system

(Braun et al., 2021)? Moreover, this reliance could influence the trust dynamic between the doctor and the patient. If the doctor can only verify and validate AI-generated results to a limited extent and is unable to communicate the basis of these results effectively, their perceived competence in the eyes of the patient might be compromised.

Additionally, the extent to which the use of AI aligns with the doctor's commitment to act in the patient's best interest becomes a pertinent question. However, the system occupies an intermediate position. The doctor retains the ability to evaluate the basic diagnostic functionality of the system, such as detecting Alzheimer's dementia via brain scans, and can assess the plausibility of AI predictions in the context of the patient's overall presentation and condition. Finally, the influence of AI on the trust relationship between doctor and patient depends significantly on the patient's individual preferences and whether they are generally open to or critical of the use of AI. In conclusion, as neither doctor nor patient can fully assess the reliability of the system, the perception of the potential reliability of the AI is founded in the doctor's trustworthiness. Here, we see how the epistemic operations for asserting the reliability of the AI come into conflict with asserting the trustworthiness of the doctor. Trust involves vulnerability to another person, as it relies on their willingness to act in alignment with one's expectations. In this context, however, the realization of these expectations depends on the performance of a system whose reliability cannot be fully evaluated by neither the individual patient nor the doctor. Similar to a potential responsibility gap (Lang et al., 2023), we see a *reliability gap* here.

This assessment underlines the advantage of addressing the epistemological foundations of trust and reliability attribution. While patients are generally able to make an informed decision regarding their doctor's trustworthiness in light of their displayed expertise in medical practice and their commitment to acting in the patient's best interest, the incorporation of an AI system can influence trust within the doctor-patient relationship. This challenge cannot be met, by adhering to the principles of trustworthy AI as proposed by the HLEGoAI.

In the third use case, we explore the deployment of chatbots for psychotherapy. These chatbots, are already in use in various forms and certified as medical products (Carrington, 2023). Typically, AI-driven chatbots are employed within the framework of blended treatment, where traditional psychotherapy is augmented by AI applications (Ehrt-Schäfer et al., 2023). Additionally, there are systems designed to provide preliminary symptom assessments, which initially occur without medical oversight. These chatbots are grounded in well-established principles of various psychotherapeutic approaches. Beyond symptom clarification, they facilitate strategies for managing everyday challenges and the long-term treatment of psychiatric symptoms. Our primary focus is on scenarios where a medical diagnosis has been established, and an AI-based chatbot has been accepted as a psychotherapeutic intervention. Such systems exhibit low determinism and are adaptive, yet their mechanisms lack transparency and cannot be fully explained. Consequently, these systems are characterized by a high degree of autonomy. Operating largely independent of the attending doctor, they are classified as replacements in terms of human-machine interaction.

Regarding the reliability of these systems, it is evident that their responses to the patient input must be generally plausible, i.e. they relate to the patient's life situation and are suitable responses to their input. However, verifying and validating their outcomes presents significant challenges. The efficacy of psychotherapy, being inherently process-oriented, cannot be easily quantified based on the suitability of individual chatbot responses. This makes assessing the validity of the underlying algorithm equally challenging. Thus, the first criterion to assess the reliability of an AI system cannot confidently be met. Since, psychotherapy is highly individualistic, the success of the system with one patient cannot reliably predict its effectiveness with others (Barron, 2021). Hence, in this use case robustness is similarly difficult to evaluate. Therefore, assessments of robustness rely heavily on the system's development history and extensive testing. Unlike the first use case, where measurable outcomes such as melanoma detection provide clear criteria for evaluation, psychotherapy outcomes are highly individualized as well. This makes it difficult to exclude risks such as overfitting. Of the criteria used to assess system reliability, only criterion (3)—plausibility—can be confidently met, and even then, its applicability to other patients remains limited. Because doctors cannot independently verify the reliability of the chatbot, they must place trust in the broader socio-technical system in which the AI has been tested and approved.

What implications does this have for the patient's trust in their doctor? If we consider a scenario where a chatbot is recommended to a patient already undergoing psychotherapeutic treatment, the acceptance of the decision to use such a tool is deeply rooted in the patient's trust in their doctor. Ideally, the chatbot is perceived as a valuable supplement to conventional therapy (Szalai, 2021). However, a potential dilemma arises: if the chatbot proves inadequate, it may negatively affect the perceived competence of the doctor; conversely, if the chatbot is received so positively that it is favored over doctor-led psychotherapy, it could undermine the central role of the doctor in the therapeutic process. Empirical research supports this tension. For instance, Ayers et al. (2023) found that chatbot responses were sometimes preferred over those of doctors, with participants rating chatbot answers superior not only in informational quality but also in perceived empathy. While the preference for a chatbot over a psychiatrist is not inherently problematic, this scenario illustrates the extent to which AI use can undermine the trust-based relationship between doctor and patient—also if the AI is perceived as effective from the patient's perspective.

From the patient's perspective, assessing the chatbot's reliability based on previously outlined criteria is nearly impossible. Patients typically lack the requisite medical expertise to evaluate the appropriateness of the chatbot's responses within the framework of psychotherapeutic treatment. Consequently, the reliability of the AI system is not perceived as intrinsic to the system itself but is instead tied to the trustworthiness of the doctor. For example, the successful functioning of the chatbot may enhance the doctor's perceived competence and reflect benevolence toward the patient – even if the patient prefers the chatbot in the end. Conversely, if the chatbot fails to perform reliably, this failure is likely to be interpreted not as a fault of the system but as a deficiency in the doctor's competence and oversight, thereby damaging the trust relationship. While the patient's trust in the doctor may be well-placed, a malfunctioning AI system could subsequently erode this trust. There is a danger that trust will not only erode, but turn into distrust, which describes not merely the absence of trust, but the expectation of harm, failure, or outcomes contrary to one's interests (Jacovi et al., 2021; Tallant, 2017). This issue is further complicated by the fact that doctors often cannot explain how AI systems function or assess their reliability, especially in light of AI's black-box nature. As a result, doctors must also place their trust in the broader socio-technical system responsible for developing and implementing the AI. This dynamic reveals the limitations of the concept of "trustworthy AI," wherein the perceived trustworthiness of AI-usage depends on its external validation within a socio-technical system. These limitations are especially apparent in healthcare, where key actors, such as doctors, are unable to directly evaluate the reliability of the AI systems they rely upon. This analysis highlights a fundamental tension between reliability and trust. Both doctors and patients are placed in a vulnerable position, as they depend on the proper functioning of the socio-technical system in which the AI operates. However, the potential betrayal of trust is not attributable to any human actor but hinges on the functionality of the AI itself, hence creating a *reliability gap*. This underscores the critical role of examining the epistemological foundations of reliability and trust, which reveals the inherent conflict and interdependence between these concepts in the context of AI in healthcare.

Conclusion

We have explored the epistemological and ethical dimensions of trust and reliability within the context of AI technologies in healthcare. Our analysis of different trust conceptions and epistemic asymmetries within the doctor-patient relationship challenges a one-size-fits-all approach to trust, demonstrating that trust relationships are shaped by fine-grained epistemic preconditions that vary significantly across different AI applications. These epistemic preconditions are particularly relevant in the doctor-patient relationship, where trust plays a foundational role in fostering communication, empathy, and shared decision-making. Specifically, the application of AI in the doctor-patient relationship exposes what we have termed a "reliability gap"—a conceptual space where the opaque nature of advanced AI systems prevents both doctors and patients from independently verifying its reliability.

This reliability gap is particularly pronounced in cases where the AI's functionality is opaque, autonomous, and applied in high-stakes scenarios, such as predictive diagnostics or therapeutic chatbots. While trust in human relationships involves affective and normative dimensions grounded in perceived competence, benevolence, and shared vulnerability, trust in the use of AI systems relies predominantly on external factors within a socio-technical context such as institutional oversight and legislative frameworks. This creates a dynamic where the trustworthiness of the AI's performance is increasingly mediated by the doctor as a proxy, whose perceived competence and commitment to patient welfare must bridge this gap. Such trust is structurally indispensable whenever patients lack direct access to AI systems. Yet it becomes particularly burdensome when opacity extends to physicians themselves, forcing both patients and doctors to rely on socio-technical guarantees that neither can independently verify.

Our use cases demonstrate that the more autonomous and opaque AI systems are, the more trust in the doctor becomes essential for bridging the reliability gap. Reliance on these technologies shifts from cognitive evaluation of reliability to affective trust in the doctor. This shift raises ethical and practical concerns, as it complicates the doctor's dual role: they must remain a trustworthy advocate for the patient while relying on systems whose reliability they themselves may not fully understand or assess. Here, trust and reliance come into conflict, as patients depend on doctors to validate the reliability of tools that exceed their epistemic grasp. How this dual reliance complicates the doctor-patient relationship in practice, at what threshold the mere absence of trust develops into active distrust and how such reliability gaps can be bridged without overburdening doctors and healthcare workers should be subject to further empirical and conceptual research.

In conclusion, the introduction of AI into healthcare necessitates a re-evaluation of trust dynamics. By highlighting the reliability gap and its implications, this paper underscores the importance of aligning epistemic preconditions for reliance, trust and trustworthiness with the ethical imperatives of patient-centred care.

Acknowledgements We thank Christin Hempeler for valuable comments and input.

Author contributions E.B. coordinated the research, drafted the discussion, developed the use cases and prepared the tables. All authors contributed substantially to the manuscript via writing, reviewing and editing.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors were partially funded by the German Federal Ministry of Education and Research (01GP2216A) and the Lower Saxony Ministry of Science and Culture (MWK) (11-76251-1251/2023 ZN 4109).

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

Declarations

Competing interests The authors have no competing financial interests to declare.

References

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., & Madai, V. I. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLoS Digital Health, 1*(2), e0000016. https://doi.org/10.1371/journal.pdig.0000016. & on behalf of the, Z. I. i.

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians [Viewpoint]. *Journal of Medical Internet Research*, *22*(6), e15154. https://doi.org/10.2196/15154

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine, 183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

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

Baier, A. (1991). Two lectures on trust: lecture 1, trust and its vulnerabilities and lecture 2, sustaining trust. In G. B. Peterson (Ed.), *Tanner lectures on human values* (Vol. 13, pp. 109–174). University of Utah.

Barron, D. S. (2021). Commentary: The ethical challenges of machine learning in psychiatry: A focus on data, diagnosis, and treatment. *Psychological Medicine*, *51*(15), 2522–2524. https://doi.org/10.1017/S0033291721001008

Blanco, S. (2025a). Human trust in AI: A relationship beyond reliance. *AI and Ethics*, *5*(4), 4167–4180. https://doi.org/10.1007/s43681-025-00690-z

Blanco, S. (2025b). *Trusting as a moral act: Trustworthy AI and responsibility*. Eberhard Karls Universität Tübingen].

Braun, M., Bleher, H., & Hummel, P. (2021). A leap of faith: Is there a formula for trustworthy AI? *Hastings Center Report*, *51*(3), 17–22. https://doi.org/10.1002/hast.1207

Bryson, J. (2018). 11–13). AI & Global Governance: No One Should Trust AI. *UNU-CPR (blog)*. https://unu.edu/cpr/blog-post/ai-global-governance-no-one-should-trust-ai

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Carrington, B. (2023). AI mental health chatbot that predicts disorders becomes first in world to gain Class IIa UKCA medical device status. https://www.limbic.ai/blog/class-ii-a

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology*, *14*(1), 53–60. https://doi.org/10.1007/s10676-011-9279-1

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT Press.

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 of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329–335. https://doi.org/10.1136/medethics-2020-106820

Durán, J. M., & Pozzi, G. (2025). Trust and trustworthiness in AI. *Philosophy & Technology, 38*(1), 16. https://doi.org/10.1007/s13347-025-00843-2

Dyrba, M., Grothe, M. J., Mohammadi, A., Binder, H., Kirste, T., & Teipel, S. J. (2018). Comparison of different hypotheses regarding the spread of Alzheimer's disease using Markov random fields and multimodal imaging. *Journal of Alzheimer's Disease, 65*(3), 731–746. https://doi.org/10.3233/jad-161197

Dyrba, M., Hanzig, M., Altenstein, S., Bader, S., Ballarini, T., Brosseron, F., Buerger, K., Cantré, D., Dechent, P., Dobisch, L., Düzel, E., Ewers, M., Fliessbach, K., Glanz, W., Haynes, J. D., Heneka, M. T., Janowitz, D., Keles, D. B., & Kilimann, I. (2021). Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: Evaluation in Alzheimer's disease. *Alzheimer's Research & Therapy, 13*(1), 191. https://doi.org/10.1186/s13195-021-00924-2. … for the Adni, A. D. s. g.

Ehrt-Schäfer, Y., Rusmir, M., Vetter, J., Seifritz, E., Müller, M., & Kleim, B. (2023). Feasibility, Adherence, and effectiveness of blended psychotherapy for severe mental illnesses: Scoping review [Review]. *JMIR Ment Health*, *10*, e43882. https://doi.org/10.2196/43882

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Faulkner, P. (2011). *Knowledge on trust*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199589784.001.0001

Faulkner, P. (2017). The Problem of Trust. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 109–128). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198732549.003.0007

Ferrario, A., & Loi, M. (2022). *How Explainability Contributes to Trust in AI* Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea. https://doi.org/10.1145/3531146.3533202

Greenhalgh, J. (2009). The applications of pros in clinical practice: What are they, do they work, and why? *Quality Of Life Research, 18*(1), 115–123. https://doi.org/10.1007/s11136-008-9430-6

Grote, T., & Berens, P. (2019). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics, 46*(3), 205–211. https://doi.org/10.1136/medethics-2019-105586

Gunkel, D. J. (2023). *Person, Thing, robot: A moral and legal ontology for the 21st century and beyond*. The MIT Press. https://doi.org/10.7551/mitpress/14983.001.0001

Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in physicians and medical institutions: What is it, can it be measured, and does it matter? *Milbank Quarterly*, *79*(4), 613–639. https://doi.org/10.1111/1468-0009.00223

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation. http://www.jstor.org/stable/10.7758/9781610442718

Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics*, *46*(7), 478–481. https://doi.org/10.1136/medethics-2019-105935

Hawley, K. (2014). Trust, distrust and commitment. *Noûs*, *48*(1), 1–20. https://doi.org/10.1111/nous.12000

Hawley, K. (2017). Trustworthy groups and organizations. In P. Faulkner, & T. Simpson (Eds.), *The philosophy of trust* (Vol. 0pp.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198732549.003.0014

HLEGoAI (2019). *Ethics guidelines for trustworthy AI.* Brussels.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63–76. https://doi.org/10.1080/00048409412345881

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). *Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada. https://doi.org/10.1145/3442188.3445923

Jones, K. (1996). Trust as an affective attitude. *Ethics*, *107*(1), 4–25. http://www.jstor.org/stable/2382241

Jones, K. (1999). Second-hand moral knowledge. *Journal of Philosophy, 96*(2), 55.

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

Jones, K. (2019). Trust, distrust, and affective looping. *Philosophical Studies*, *176*(4), 955–968. https://doi.org/10.1007/s11098-018-1221-5

Keren, A. (2007). Epistemic authority, testimony and the transmission of knowledge. *Episteme, 4*(3), 368–381. https://doi.org/10.3366/E1742360007000147

Lang, B. H., Nyholm, S., & Blumenthal-Barby, J. (2023). Responsibility gaps and black box healthcare AI: Shared responsibilization as a solution. *Digital Society*, *2*(3), 52. https://doi.org/10.1007/s44206-023-00073-z

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report, 49*(1), 15–21. https://doi.org/10.1002/hast.973

Lorenzini, G., Arbelaez Ossa, L., Shaw, D. M., & Elger, B. S. (2023). Artificial intelligence and the doctor–patient relationship expanding the paradigm of shared decision making. *Bioethics*, *37*(5), 424–429. https://doi.org/10.1111/bioe.13158

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 3–25). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00001-0

McLeod, C. (2023). Trust. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*.

Mökander, J., Sheth, M., Watson, D. S., & Floridi, L. (2023). The switch, the ladder, and the matrix: Models for classifying AI systems. *Minds and Machines, 33*(1), 221–248. https://doi.org/10.1007/s11023-022-09620-y

Montgomery, K. (2005). *How Doctors think: Clinical judgment and the practice of medicine*. Oxford University Press. https://doi.org/10.1093/oso/9780195187120.001.0001

Nouis, S. C. E., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision- making: A qualitative study of healthcare professionals' perspectives in the UK. *BMC Medical Ethics, 26*(1), Article 89. https://doi.org/10.1186/s12910-025-01243-z

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Onder, O. (2022). Epistemolojik ve Etik Acidan Klinik Karar Destek Sistemleri. In T. Bardakçı & M. I. Karaman (Eds.), *Yapay Zeka Etiği* (pp. 147-160). Isar Yayınları.

Onder, O. (2025). Klinik Karar Destek Sistemleri Bağlamında Tıpta Yapay Zeka Kullanımında Etik Sorun Alanları (Publication No. 949041) [Doctoral dissertation, Istanbul University] YÖK Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/

Pasquale, F. (2015). *The black box society*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Patel, V. L., Arocha, J. F., & Zhang, J. (2012). 736 medical reasoning and thinking. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0037

Polanyi, M. (1966). *The Tacit dimension*. Chicago University Press.

Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, *3*(3), 735–744. https://doi.org/10.1007/s43681-022-00200-5

Riedl, R., Hogeterp, S. A., & Reuter, M. (2024). Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice [Original Research]. *Frontiers in psychology*, *Volume 15–2024*. https://doi.org/10.3389/fpsyg.2024.1422177

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. pearson.

Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749–2767. https://doi.org/10.1007/s11948-020-00228-y

Safdari, A. (2025). Toward an empathy-based trust in human-otheroid relations. *AI & SOCIETY*, *40*(5), 3123–3138. https://doi.org/10.1007/s00146-024-02155-z

Sauerbrei, A., Kerasidou, A., Lucivero, F., & Hallowell, N. (2023). The impact of artificial intelligence on the person-centred, doctor-patient relationship: Some problems and solutions. *BMC Medical Informatics and Decision Making, 23*(1), 73. https://doi.org/10.1186/s12911-023-02162-y

Shevtsova, D., Ahmed, A., Boot, I. W. A., Sanges, C., Hudecek, M., Jacobs, J. J. L., Hort, S., & Vrijhoef, H. J. M. (2024). Trust in and acceptance of artificial intelligence applications in medicine: Mixed methods study. *JMIR Human Factors, 11*, Article e47031. https://doi.org/10.2196/47031

Simmler, M., & Frischknecht, R. (2021). A taxonomy of human–machine collaboration: capturing automation and technical autonomy. *AI & SOCIETY*, *36*. https://doi.org/10.1007/s00146-020-01004-z

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack, C. R., Jr., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., … Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement, 7*(3), 280–292. https://doi.org/10.1016/j.jalz.2011.03.003

Stroud, A. M., Minteer, S. A., Zhu, X., Ridgeway, J. L., Miller, J. E., & Barry, B. A. (2025). Patient information needs for transparent and trustworthy cardiovascular artificial intelligence: A qualitative study. *PLoS Digital Health, 4*(4), e0000826. https://doi.org/10.1371/journal.pdig.0000826

Szalai, J. (2021). The potential use of artificial intelligence in the therapy of borderline personality disorder. *Journal of Evaluation in Clinical Practice*, *27*(3), 491–496. https://doi.org/10.1111/jep.13530

Taddeo, M. (2010a). 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. (2010b). Trust in technology: A distinctive and a problematic relation. *Knowledge Technology & Policy*, *23*(3), 283–286. https://doi.org/10.1007/s12130-010-9113-9

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

Tallant, J. (2017). Commitment in cases of trust and distrust. *Thought, A Journal of Philosophy, 6*(4), 261–267. https://doi.org/10.1002/tht3.259

Topol, E. J. (2019). *Deep medicine: How artificial intelligence can make healthcare human again* (1st ed.). ed.). Basic Books.

Ursin, F., Lindner, F., Ropinski, T., Salloch, S., & Timmermann, C. (2023). Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? *Ethik in der Medizin, 6*(5), 52138. https://doi.org/10.1007/s00481-023-00761-x

Ursin, F., Timmermann, C., & Steger, F. (2021). Ethical Implications of Alzheimer's Disease Prediction in Asymptomatic Individuals through Artificial Intelligence. *Diagnostics (Basel, Switzerland), 11*(3). https://doi.org/10.3390/diagnostics11030440

Vernon, D. (2014). *Artificial cognitive systems: A primer*. The MIT Press. http://www.jstor.org/stable/j.ctt17kk720

Walker, M. (2006). Moral repair: Reconstructing moral relations after wrongdoing. 1–250. https://doi.org/10.1017/CBO9780511618024

Willem, T., Krammer, S., Böhm, A. S., French, L. E., Hartmann, D., Lasser, T., & Buyx, A. (2022). Risks and benefits of dermatological machine learning health care applications—An overview and ethical analysis. *Journal of The European Academy of Dermatology and Venereology, 36*(9), 1660–1668. https://doi.org/10.1111/jdv.18192

Winkler, J. K., Sies, K., Fink, C., Toberer, F., Enk, A., Abassi, M. S., Fuchs, T., & Haenssle, H. A. (2021). Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *European Journal of Cancer*, *145*, 146–154. https://doi.org/10.1016/j.ejca.2020.12.010

Wu, Q., Jin, Z., & Wang, P. (2022). The relationship between the physician-patient relationship, physician empathy, and patient trust. *Journal of General Internal Medicine, 37*(6), 1388–1393. https://doi.org/10.1007/s11606-021-07008-9

Zondag, A. G. M., Rozestraten, R., Grimmelikhuijsen, S. G., Jongsma, K. R., van Solinge, W. W., Bots, M. L., Vernooij, R. W. M., & Haitjema, S. (2024). The effect of artificial intelligence on patient-physician trust: Cross-sectional vignette study. *Journal of Medical Internet Research, 26*, e50853. https://doi.org/10.2196/50853

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.