Minds and Machines (2026) 36:5
https://doi.org/10.1007/s11023-025-09754-9

Being Pragmatic About Reliance and Trust in Artificial Intelligence

Andrea Ferrario[1,2,3]

Received: 21 February 2025 / Accepted: 5 November 2025

Abstract

The ongoing debate about reliance and trust in artificial intelligence (AI) systems seems to be never ending, challenging our understanding and application of these concepts in human-AI interactions. In this work, we argue for a pragmatic approach to solve this conundrum by focusing on reliance and the three key expectations that should guide human-AI interactions: appropriate reliance, efficiency, and motivation by objective reasons. By focusing on these expectations, we show that it is possible to reconcile reliance with trust in a manner that is both theoretically sound and practically useful. As it turns out, reliance is the key relation of interest while trust in AI is a derived concept that helps explaining these expectations. Our reliance-centered framework does not dismiss the concept of trust in AI but repositions it as a key property of reliance, offering a pragmatic alternative to classical rational or motivational accounts of trust that prove difficult to apply in the context of human-AI interactions. As AI continues to integrate into society, particularly in high-stakes environments like healthcare, our pragmatic approach provides a practical and meaningful framework for addressing the nuances of trust in AI.

Keywords Reliance · Trust · Artificial Intelligence · AI · Explainable AI

✉ Andrea Ferrario
andrea.ferrario@ibme.uzh.ch

[1] Institute of Biomedical Ethics and History of Medicine, University of Zurich, Zurich, Switzerland

[2] University of Applied Sciences and Arts of Southern Switzerland (SUPSI), Dalle Molle Institute for Artificial Intelligence (IDSIA), Lugano, Switzerland

[3] ETH Zurich, Zurich, Switzerland

Published online: 15 December 2025

1 Introduction

As we approach the end of 2025, it is somewhat disheartening to note that we are still engaged in discussions about reliance and trust in artificial intelligence (AI) systems. Depending on one's perspective, this debate might be seen as a testament to the intellectually engaging complexity of these issues, or as a sign of our inability to find definitive answers. For sure, the debate on reliance versus trust has a long history in philosophy, focusing originally on interpersonal relations and converging on the idea that, somehow, 'trust is reliance plus an extra factor' (Baier, 1986; Pettit, 1995; Dasgupta, 1988; Cogley, 2012; Hawley, 2019). Over time, the distinction between reliance and trust has become increasingly pertinent in the context of AI as well. Many authors have endorsed the possibility of trusting AI systems, actively discussing what makes these systems worthy of our trust (Floridi et al., 2018; Floridi, 2020; Coeckelbergh, 2012). A few scholars argue that when humans engage with AI, reliance is the only feasible relation with these systems instead (Ryan, 2020; Hatherley, 2020). This view typically suggests that the relation of trust, as traditionally conceived, is inapplicable to non-human entities, which lack the mental states and moral agency that, in their view, typically ground trusting. The dispute is still ongoing and seems to have no foreseeable end. While theoretical debates have yielded no consensus on the concepts of reliance and trust in AI, empirical studies investigating human-AI interactions have similarly fallen short, failing to produce a consistent and methodologically robust body of empirical evidence on the dynamics of trusting AI, as systematic reviews have highlighted (Vereschak et al., 2021; Benk et al., 2024).

In this work, we show that it is possible to conceptualize 'trust in AI' without reducing it to mere reliance on AI systems. To do so, we challenge the 'trust is reliance with an extra factor' paradigm proposing a pragmatic approach to reconciling reliance and trust in AI systems, grounded in a shift within human-AI interaction research from a focus on trust to an emphasis on reliance (Schemmer et al., 2023). Namely, we treat reliance as the more fundamental and tractable construct, using it as the foundation of our approach on human-AI interactions. Doing so, we move beyond the traditional discourse typically positioning the two concepts in opposition to one another, while addressing the theoretical issues that affect empirical research on trust in AI. The pragmatic perspective we advocate begins by addressing a key question: '*What do we expect from our relying on AI systems?*' Naturally, expectations can vary widely depending on the specific context and application. However, we argue that there exists a minimal set of expectations that should consistently guide our interactions with AI. These include (1) appropriateness, (2) efficiency, and (3) calibration, that is, reliance grounded in the system's objective capabilities.

First, we show that appropriate reliance is the first objective of human-AI interactions; for instance, within the context of AI supporting decision-making, it can be clearly formalized in terms of the performance of human-AI teams (Schemmer et al., 2023). It does not provide insights on trusting in that relation, but contribute to trusting for future interactions. In particular, we address the efficiency of reliance by endorsing the account of trust that conceptualizes it as a quantitative *property* of reliance, rather than a *relation* traditionally positioned in contrast to reliance (Taddeo, 2010). We show that efficiency of appropriate reliance relations is equivalent to hav-

ing high levels of trust in the AI due to either the optimization of invested resources, called 'monitoring,' or a strong confidence in the reliability of the system. Thus, in our pragmatic account, trust quantifies the efficiency of our appropriate reliance on the AI's outputs. As a result, whenever there is reliance, there is also trust. We provide an explicit characterization of the levels of trust in a reliance relation with an AI system considering the investment of monitoring resources and the expected gains in the interaction with the machine (Taddeo, 2010; Loi et al., 2023). Finally, we clarify the relation between reliance, trust, and trustworthiness of AI, showing that 'calibrating reliance with the trustworthiness of AI' leads to appropriate, efficient reliance caused by the objective properties that make the AI worthy of trust. Our account aligns with the 'Trustworthy AI' framework (Floridi et al., 2018) and it promotes the idea that the trustworthiness of AI is the composite capability of AI systems that gives reasons to *rely on* them with *high levels of trust*.

Our pragmatic approach separates trust in AI from expectations of competence, goodwill, moral attributes, anchoring the concept in the longstanding tradition that sees it antithetical to monitoring (Baier, 1986; Dasgupta, 1988; Keren, 2014). This choice provides a degree of protection against the limitations that affect the widespread accounts of trust that focus on the rationality of the interaction or the trustor's motives (Zanotti et al., 2023). Further, it grounds the characterization of trust as a quantitative property of reliance, as proposed by various authors (Taddeo, 2010; Loi et al., 2023), in the criterion of efficiency in successful instances of human-AI reliance relations. The measurability of appropriate and efficient reliance, as well as the use of calibration offer a framework that can be implemented to test hypotheses on trust in AI in empirical studies. Thus, our approach has the potential to address the lack of theoretical coherence affecting empirical studies on trust in AI (Benk et al., 2024) and to stimulate debate on whether the attribution of human-like qualities to AI is necessary when discussing trust in this technology. This said, our account does not dare to find a solution to the dispute on trust in AI that would garner universal agreement. We believe that there will always be scholars in favor of a characterization of trust in AI that is reminiscent of what the concept seems to satisfy in certain relations involving humans, starting with mental states or moral obligations, affective commitments, and different virtues–even more considering the persuasive capabilities of large language model-based applications (Floridi & Nobre, 2024; Shanahan et al., 2023). Nonetheless, our pragmatic perspective on reliance and trust in AI suggests that we should prepare ourselves for the possibility that trust may play a less central role in human-AI interactions than as previously thought (or hoped). This despite its importance in interpersonal relations and for society as argued in a vast body of literature. Or, at least, that eventually trust has to pass the torch to reliance, which may prove simply more relevant and practical in the context of human-AI interactions.

2 An Age-Old Struggle: Reliance vs. Trust

In this Section, we provide an overview of two key constructs: reliance and trust. We then discuss how the concepts are treated in the context of human-AI interactions. This will provide us with an overview of the problems affecting 'trust in AI.'

2.1 Reliance

We rely on something happening or on some entity doing something (Holton, 1994). For instance, we rely on the ticket seller to process our ticket at the counter so we can catch our train, on the microwave door to open after two minutes so we can eat, and on the tide to reach its peak at 08:30 a.m. so we can begin sailing. Thus, we rely on people, artefacts and natural phenomena alike. By relying on these entitites, "we act in a way that is premised on their being of a certain character or on their being likely to act under various circumstances in a certain way" (Pettit, 2004, pag. 109). More formally, Goldberg states: "[given that] X is a person, artifact, or natural process, and φ is an action, behavior or process, to rely on X to φ is to act on the supposition that X will φ" (Goldberg, 2020, pag. 97). Here, the term 'supposition' captures both the dependence on X to φ and the expectation that X will indeed φ.[1] However, in general, this supposition in relying on does neither require X being aware to be relied on, nor us holding beliefs about whether X will φ. In fact, reliance is compatible also with the absence of belief, including without being aware of relying at all (Goldberg, 2020).[2]

2.2 Trust

While reliance is an uncontroversial notion, capturing our dependence on the world and our actions based on that dependence, trust is more difficult to define. There is something about trust that differentiates it from reliance and philosophers have been trying to describe it since long time. Structurally, reliance and trust are traditionally conceptualized as relations, with reliance being necessary for trust but distinct from it (Hawley, 2014; Baier, 1986; Pettit, 1995). Qualitatively, reliance is often considered a simpler, less intriguing relation than trust, as emphasized by comparisons, such as 'mere reliance' versus 'genuine trust' commonly used to differentiate the two in philosophical discussions (Hawley, 2014; Zanotti et al., 2023; Durán & Pozzi, 2025). This prevailing view suggests that trusting requires reliance,[3] with trust being a special form of reliance, as described by the celebrated phrase "trust is standardly thought to involve reliance, plus some extra factor" (Hawley, 2014, p. 5). For instance, rational accounts—also called 'epistemic accounts' (Zanotti et al., 2023)—describe this

[1] Holton offers a characterization of this supposition stating that "[w]hen I rely on something happening I need not believe that it will happen. But [...] I need to work around the supposition that it will" (Holton, 1994, pag. 3). However, "[t]o rely on a person doing something is not just to rely on a certain state of affairs happening: the state of affairs in which they do that thing [...] I plan [to rely] on the supposition that you will stay there because you are motivated to stay there, and not just because you have no choice" (Holton, 1994, pag. 3). Further, the dependence in relying is explicitly reflected in the Cambridge Online Dictionary definition of reliance as "the state of depending on [...] something or someone"—available at https://dictionary.cambridge.org/dictionary/english/reliance.

[2] In this work, we focus on practical forms of reliance, specifically where reliance is goal oriented. This means acting on the supposition that X will φ to achieve a specific goal. This approach parallels the idea of trusting someone to perform a particular task to achieve an objective, rather than trusting someone unconditionally.

[3] Hawley is explicit on this point: "A common core of trust is practical reliance: part of trusting people to do things is an expectation that they will in fact do those things, whilst part of trusting someone's word is relying upon what she says. Trust involves reliance [...]" (Hawley, 2019 pag. 2).

extra factor by emphasizing the importance of beliefs and rational choices over the trustor's motives as the key determinant of trust (Gambetta, 2000; Mollering, 2006). In particular, Hardin (1992) states that to trust is a species of reliance grounded in a belief that reflects a relationship between the interests of the trustor and the trustee–an account known as 'encapsulated interest.' Affective and normative accounts—also known as 'motivational accounts' (Zanotti et al., 2023)—load trust with expectations of goodwill (Baier, 1986; Cogley, 2012), interests (Hardin, 2002), moral commitments, or obligations on the trustee instead (Nickel, 2007). As famously argued by Baier, "trust [...] is accepted vulnerability to another's possible but not expected ill will (or lack of good will) towards one" (Baier, 1986, p. 235). In other words, trust is reliance on another's good will (Baier, 1986).

While these approaches capture relevant aspects of that we commonly expect from the concept of 'trust,' they are prone to counterexamples, which ultimately impede a univocal characterization of trust as an interpersonal *relation*. For instance, rational accounts of trust may not adequately characterize emotionally charged situations where trust is expressed without a clear rational basis. Consider a child who trusts their parent to protect them in a dangerous situation, not because of any rational assessment of the parent's capabilities, but because of an innate emotional bond and their limited understanding of the world. Motivational accounts may fall short in contexts where rationality plays a major role in decision-making instead. For instance, in professional environments, trust can be grounded in the rational evaluation of people's competence and reliability, with little regard to affective or moral factors. In particular, Holton's famous 'confidence trickster' example suggests that reliance on good will does not explain trusting (Holton, 1994). In summary, the different accounts of 'trust is reliance plus an extra factor' seem to fail to encompass the full variability of trusting across different contexts and interpersonal interactions. Simply put, they work well in some paradigmatic instances, but are prone to counterexamples.[4]

2.3 Reliance and Trust in Human-AI Interactions

Reliance in AI. Reliance and trust have also entered the field of human-AI interaction, gaining prominence with the growing adoption of machine learning models in modern AI systems. As discussed in Sect. 2.1, the concept of reliance applies uncontroversially to interactions with artefacts; in particular, we rely on AI systems to achieve different goals. For instance, in the case of decision-making, we rely on an AI by delegating to the machine the task of computing an accurate prediction of a phenomenon of interest, such as a numerical score (e.g., the expected temperature of a device) or a label (e.g., the classification of a medical images) and use it for making decisions as opposed to not using them—that is, rely on oneself or on colleagues, social or traditional media, or other technologies (e.g., ChatGPT).[5] The framework

[4] The interested reader may further explore the literature on reliance and trust through the monograph by Simon (2020) and the entry on 'Trust' in the Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/entries/trust/).

[5] Using Goldberg (2020)'s definition of reliance again, X is the AI system, and φ is to compute a prediction that is used for decision-making.

of reliance on AI versus self-reliance is central to the literature on human-AI interactions, foundational in medical AI research, and widely explored in the philosophy of AI (Schemmer et al., 2023; Esteva et al., 2017; Topol, 2019; Bjerring & Busch, 2021). In fact, for the purposes of empirical studies in human-AI interaction, self-reliance is typically chosen as the contrasting case to relying on an AI system, as it provides a tractable baseline for measurement. This type of human-AI interactions—especially within high-risk applications such as healthcare—represents contexts where reliance and trust in AI are especially important, and collaboration between humans and AI is required (Lee & See, 2004). This said, we also rely on AI systems for various other purposes, such as entertainment, creative expression, social interaction, and support for life needs, including mental health. In these contexts, much like in decision-making, self-reliance serves as an alternative to relying on the system to provide entertainment, creative ideas, social engagement, or well-being support. LLM-based apps, such as ChatGPT or Claude, are pretty good as providing us with these types of assistance. Finally, in human-AI interactions, *reliability* designates what makes an AI system worthy of being relied upon. Technically, it refers to the system's capacity to perform its intended function without failure, typically modeled as a probability over time. Philosophical debates focus on how such reliability can be justified epistemically, with proposals currently centering on transparency and computational reliabilism (Durán & Pozzi, 2025).

Trust in AI. As in the case of interpersonal relations—see Sect. 2.2—problems arise when trying to define the concept of 'trust in AI' while characterizing the differences between relying and trusting these systems. Some scholars debate whether trusting AI is even possible. Among those who argue positively, Lee and See's in their seminal work 'Trust in Automation: Designing for Appropriate Reliance' (Lee & See, 2004), while focusing on the type of AI called automation, are clear on the possibility of trusting this technology. In different works, Floridi and colleagues have similarly endorsed the notion that AI can be trusted, discussing the conditions under which AI technology becomes worthy of trust (Floridi et al., 2018; Floridi, 2018, 2020). Further, Bjerring and Busch discuss the role of trust in 'black-box' medicine, namely, medical practice mediated by AI systems (Bjerring & Busch, 2021), while Alvarado argues that epistemic trust is the only legitimate kind of trust to allocate to AI technology (Alvarado, 2023). However, other authors argue that while AI systems meet the requirements for reliance, they lack the capacity of being trusted (Ryan, 2020; Hatherley, 2020; DeCamp & Tilburt, 2019). The typical argument goes as follows: our understanding of trusting comes from interpersonal relations, where trust is typically characterized as reliance plus an extra factor, see Sect. 2.2. Depending on the chosen account, trust in interpersonal relations requires a trustor with specific mental states, motivations, or adherence to moral obligations. Then, if we translate trust into the context of human-AI interactions, as AI systems do not possess these features, our accounts of trust are not applicable to them (Ryan, 2020; Hatherley, 2020; DeCamp & Tilburt, 2019). All that remains is 'mere' reliance.

What is still open for debate on trust in AI? A remark is due. Trust undoubtedly originated within the context of relationships between humans—including their orga-

nizations—and, potentially, superhuman entities in religious settings (Fabris, 2020). However, while the concept of reliance on AI can be defined in technical terms related to the predicting capabilities of AI systems, the appropriateness of translating a complex, multidimensional concept like trust from human relations to human-AI interactions remains to be proven. After all, it is a choice. Therefore, despite its popularity, it must be justified against alternative approaches by providing reasons for its use as a framework for discussing trust in AI—including the possibility that such systems may not be trusted at all. In addition, the absence of a univocal definition of trust in interpersonal relations does not support the use of any specific characterization of trust—e.g., motivational or rational ones—as a sound starting point for examining this concept in human-AI interactions. Alternatively, one could claim that a more fundamental construct whose projection into the distinct contexts of interpersonal relations and human-machine interactions determines their possibly irreconcilable approaches to trust is more appropriate to address the problem of defining trust in AI. Proponents of this approach would reject the aforementioned argument about the impossibility of trusting AI systems by asserting the incomparability between interpersonal relations and interactions with machines. However, endorsing this perspective would require the characterization of this more fundamental trusting construct and its projection into two separate classes of agential relations in some detail.[6] In summary, there are several accounts of trust in AI that can be promoted and discussed. In addition, these accounts may focus on different types of interactions and AI systems, such as automation to support decision-making, social robots promoting socialization: see, for instance, Lee and See's account of trust in automation (Lee & See, 2004) or Coeckelbergh's 'contractarian-individualist' and 'phenomenological-social' approaches to trusting robots (Coeckelbergh, 2012). This said, each proposal has to be appropriately motivated, starting with the common 'trust is reliance plus an extra factor' one, which is often used as an argument against trusting AI technology tout court. No perspective is set in stone; all accounts require careful motivation and context-specific justification. We will follow this general recommendation as soon as we introduce our approach to trust in AI in the next section.

3 A Pragmatic Perspective on the Reliance vs. Trust Conundrum

3.1 Our Approach: Overview

We want to address the issue of conceptualizing trust in AI from a different perspective. Instead of analyzing the theoretical distinctions between reliance and trust once again, drawing inspiration from some account of trusting and its counterexamples from the ever-growing body of literature on the topic, we promote a *pragmatic* approach to solve this conundrum. By 'pragmatic' we do not mean linguistic pragmatics, but an orientation rooted in the pragmatist tradition, where concepts are

[6] Note that Grodzinsky et al. (2011) tried to introduce such an account–called **TRUST**–that projects onto human interpersonal relations, human-AI as well as AI-AI interactions. However, their 'Ur-trust' account consists only of high-level principles that have to do with delegation and risk among others.

assessed by their relation and consequences for action—see Peirce's idea of belief as action (Peirce, 1940) as well as more recent work on pragmatic ethics of AI (Coeckelbergh, 2020; Van Wynsberghe & Robbins, 2014). Our use of pragmatic signals that the approach to reliance offered here is meant to be action-guided: it evaluates reliance in terms of expectations and practical outcomes in real-world human-AI interactions. Thus, our account of reliance and trust in AI is grounded in practical considerations, rather than in different extra factors or other assumptions. Its value lies in its ability to capture key aspects of real-world interactions with AI technology in a few steps, without requiring assumptions about humanizing these systems as a basis for discussing trust in them or showing the impossibility to do so. In the remainder of this work, we aim to show that our reframing of the reliance-trust in AI problem helps avoid the conceptual pitfalls of directly applying the 'trust as reliance plus some extra factor' account to human-AI interactions, without abandoning the idea of genuinely trusting AI systems. Our pragmatic argument starts with this question: '*What do we expect from our relying on AI systems?*' We believe that, despite the diverse types of human-AI interactions and the variety of AI systems commonly used in a myriad of applications (including the more recent LLM-based services), when it comes to relying on AI there are three fundamental expectations that users consistently hold and guide their actions. First, our reliance on AI should be *appropriate*. Essentially, the AI outputs, such as predictions or recommendations, are adequate to achieve the goals motivating the reliance on the system. Second, reliance relations should be *efficient*, namely, performed by optimizing available resources (e.g., time and information) over time. Different theories, including Cognitive Load Theory (Sweller, 1988), suggest that decision-making efficiency is promoted when resources are optimally utilized. As human decision-makers (e.g., physicians) often operate under constraints such as limited information, time, and cognitive capacity, they need to rely on AI systems with an optimal resource allocation for their decision-making. Finally, our reliance on AI systems should be rooted in the objective capabilities, including predictive performance, safety, robustness, and transparency of these machines. If that is the case, the reliance is *calibrated*. In modern terms, these capabilities constitute what people commonly refer to as the trustworthiness of the AI system. Note the use of both reliance and trust(worthiness) for calibration: we will clarify this point in what follows.

Being appropriate, efficient, and calibrated are what, we believe, people expect from relying on AI technology. Removing any one of these expectations substantially undermines the quality of reliance, which could lead to critically flawed outcomes. In fact, without appropriateness, relying would lead to sub optimal interactions or even harmful decision-making outcomes. Without efficiency, the practical utility of relying on AI would be diminished. Since human cognitive resources are limited, interactions must be streamlined. An AI system that is accurate but utilized in inefficient reliance relations, namely, requiring excessive time, effort, or cognitive resources can lead to fatigue, where users are overwhelmed and unable to maintain high quality interactions over time. Further, if reliance is not grounded in the objective capabilities of the AI system, users risk unjustified reliance or over-reliance, leading to decisions based on wrong AI's predictions, recommendations, and other outputs, or under-reliance, where useful AI outputs are ignored.

Anchoring our discourse on trust in AI from the pragmatics of relying will allow us reconciling trust in AI with the expectations we have while we interact with these systems. Namely, we will conceptualize trust in AI to satisfy the criteria of appropriateness, efficiency, and calibration of relying on these systems. Thus, differently from accounts that root trust in AI in a set of ethical principles and commitments that relate its designers, regulatory organizations, and the public, or straightly negate its conceptual legitimacy, using our pragmatic approach we will describe trust in AI as *a phenomenon of reliance* in these systems. In our account, ethical considerations and human motives allow describing the *how* and *why* of trust but are not key to define *what* trust is. We will elaborate on these points in what follows, where we discuss these three expectations of reliance and trust in AI in some detail. For clarity, in Sections 3.2-3.4, we will focus on the decision-making context and use a healthcare-related example to support our discussion. Healthcare is a central domain for studying reliance, both in philosophical debates and in human–AI interaction research. Nonetheless, our considerations extend to other forms of human-AI reliance as well: the definition of reliance (see Section 2.1) applies broadly to other forms of human-AI reliance, encompassing diverse domains, delegated actions, and goals.

3.2 Being Accurate when Relying on AI's Inputs: Appropriate Reliance

To rely on AI is acting on the supposition that these systems will provide outputs—such as predictions, recommendations—to achieve a specific goal. This goal might involve making informed decisions, choosing an action on AI-driven suggestions, among others. Thus, the first and most fundamental requirement for relying on AI is that these outputs must be adequate for achieving the intended goal. To this end, for instance, researchers in the field of human-AI interactions have started increasingly exploring how to improve human-AI team performance (Schemmer et al., 2023; Yang et al., 2020; Lai et al., 2021). In their investigations, they promote a shift of focus from trust to 'appropriate reliance,' as "human decision-makers should not simply rely on AI advice, but should be empowered to differentiate when to rely on AI advice and when to rely on their own" (Schemmer et al., 2023, p. 411).

To describe what appropriate reliance means in this context, let us consider the case of a human agent named Bob interacting with an AI system. Bob is a pulmonologist working in a bustling urban hospital and in need to ascertain whether a given pulmonary X-ray scan shows a case of COVID-19 or not using a dedicated medical AI system. During this interaction, Bob needs to choose whether to act upon the system's prediction or theirs to formulate a diagnosis and select the best treatment. At a first glance, this binary decision-making problem may seem to be an oversimplification of reality. However, it allows describing several interactions where humans are assisted by AI technology in their decision-making (Schemmer et al., 2023). For the sake of clarity, we say that the reliance relation between Bob and the AI starts when Bob decides to consult the system and generate its prediction. By relying on the AI system, Bob delegates to the machine the goal of computing an accurate prediction for the medical image he needs to classify. The reliance relation ends with Bob's decision to follow the system's prediction or not for his clinical decision-making. Then, we say that Bob's reliance on the AI is appropriate whenever he relies on a

correct system's prediction, and he relies on his prediction if the system's is incorrect (Schemmer et al., 2023). Consequently, overreliance would happen when Bob takes actions on incorrect predictions. Underreliance denotes the excessive use of Bob's self-reliance instead. Appropriate reliance on the AI may be accidental or result from Bob's 'successful appropriation' of the AI, as the system enhances his cognitive and epistemic faculties .

However, the accuracy of the prediction typically cannot be assessed during decision-making because the ground truth is not available. If it were, there would be no need for AI support. In real-world interactions, evaluating an AI's prediction accuracy requires gathering additional information, e.g., performing additional medical exams, comparing expert opinions, and analyzing historical data, as well as investing further resources. Despite this general limitation, common in decision-making contexts, appropriate reliance can serve as a tool for testing hypotheses in empirical studies using rather intuitive methodologies.[7] In summary, appropriate reliance captures the primary objective of relying with AI systems: using this technology to achieve a favourable outcome, e.g., making decisions based on accurate predictions, as the result of having correctly interacted with the system. However, it is not, by itself, sufficient to characterize what it means to trust an AI system. For that, we must probe further into the features of reliance in human-AI interactions.

3.3 Managing Resources in Human-AI Interactions: Efficient Reliance

Let us return to Bob and his medical AI. Bob is highly skilled and relies on his medical AI appropriately in his clinical practice. Over time, he has demonstrated the ability to endorse the system's predictions when they are accurate and withhold reliance when the system's performance dangerously falters, as in the instance where it incorrectly classifies an X-ray image as a case of COVID-19 infection. However, unlike his competent colleagues who utilize the same AI, Bob really takes his time to decide to rely on the AI or not. For instance, he frequently consults the AI system's official documentation, comparing the system's specifications with the latest scientific findings in the bioinformatics of COVID-19, such as (Shi et al., 2020). During clinical rounds, he often interrupts other pulmonologists, asking for second opinions and evidence supporting their decision-making. He also requested the assistance of a data engineer to create a custom interface that allows him to deep-dive on some technical functionalities of the AI system, including its retraining schedules and the metadata of the deep learning model it uses to classify images. In other words, Bob invests substantial resources—time, cognitive effort, and social capital—into each interaction with the AI system before deciding to rely on its outcomes. Other pulmonologists also rely on the AI appropriately, yet they do so with significantly less resource

[7] For instance, Schemmer et al. (2023) offer a metric measuring appropriate reliance that is based on the degrees of reliance on the AI and self-reliance of its users. Other measures are possible, giving researchers the possibility to investigate appropriate reliance empirically with an array of tools. To assess the appropriateness of relying on AI systems for enhancing creativity, providing entertainment, or offering mental health support, we could use self-assessment questionnaires to rate the quality of AI outputs, as well as behavioral measures such as the adoption rate of AI recommendations and the completion rate of assigned tasks.

investment. When Bob is jokingly confronted about his behavior, he responds with a tone of concern: 'Despite all evidence supporting the accuracy of the AI, you can never be too sure.'

This example illustrates that different human agents—Bob and his colleagues—may invest different amounts of resources and achieve the same level of appropriate reliance on the same AI system. It also suggests that, somehow, the investment of resources in an interaction with an AI system is related to what we commonly call 'trust.' That is, it seems that the efficiency (or the lack of it) in managing the resources invested while potentially relying on an AI tells us something about the trusting attitude we may have towards the system. In this Section, we elaborate on this important consideration in two steps. First, we introduce a perspective of reliance and trust that challenges the common 'trust as reliance plus an extra factor' paradigm. Then, we show that this perspective allows relating the efficiency of appropriate reliance relations with the concept of trust.

3.3.1 Step I. Trust as a Property of Reliance Relations

There exists a perspective on reliance and trust that is applicable to human-AI interactions, offering a reconciliation of these concepts without engaging in the traditional disputes that have traditionally characterized their discussion. Although this perspective is less commonly embraced than the prevalent view suggesting that trust is reliance with an extra factor, it deserves our attention, because our pragmatic approach grounds it into the efficiency of reliance on AI systems. This perspective posits that reliance is the fundamental relation we should focus on, regardless of whether it involves humans or machines, with trust being one of the quantitative properties inherent in that relation. According to this view, whenever we rely on someone or something, we always trust them, albeit to varying degrees depending on the interaction. Let us discuss this point in more detail.

To the best of our knowledge, this approach was first discussed by Taddeo, who proposed that the form of trust emerging in digital contexts—called 'e-trust'—should be understood as a *property* of certain relations within those contexts (Taddeo, 2009, 2010). E-trust, according to Taddeo, is a second-order property of a first-order relation in digital contexts which, for simplicity, we may equate with reliance. Taddeo argues that this choice has the benefit to define "e-trust as a second-order property of first-order relations, whereas e-trust is classically (if mistakenly) defined as a first-order relation that occurs between agents" (Taddeo, 2010, p. 249). Subsequently, other authors have endorsed the idea that trust can be conceptualized as a property of reliance in various domains, for instance logic, discussing the reliance on testimony and trusted communications between agents (Primiero & Taddeo, 2012; Primiero, 2020). More recently, Loi and colleagues have provided a mathematical formalization of trust as a property of reliance relations, focusing on the quantification of the intensity of trusting (Loi et al., 2023). (We will explore this approach in further detail in the following sections.)

By viewing trust as a property of reliance relations, rather than a relation itself to be compared to relying, as traditionally conceptualized, many of the traditional conflicts in the reliance vs. trust debate are rendered obsolete. For instance, this approach

bypasses the need to investigate the motives behind the trustor's trust to determine its appropriateness to specific situations, such as in the case of motivational accounts of trust. Moreover, it sidesteps the conceptual challenge of applying rational accounts of trust in contexts dominated by affective states, and vice versa. While this perspective may appear to downgrade trust to the role of being just a qualifier of reliance, we believe its grounding in the uncontroversial and broad concept of reliance is a strength. First, focusing on reliance—see Goldberg's definition (Goldberg, 2020)—enables the application of this approach to both interpersonal relations and human-AI interactions (Loi et al., 2023). In addition, this choice aligns with recent trends in the human-AI interaction domain, as discussed in Section 3.2, and, as a result, legitimizes the idea of studying trust in machines, such as AI systems, as well as its measurement in empirical studies (Benk et al., 2024).[8]

Critics might still argue that reducing trust to just a property of reliance oversimplifies the concept, failing to address its complex, multidimensional nature that encompasses affective, cognitive, and normative elements depending on the account chosen to characterize it. Further, specifying that trust is a property of reliance without discussing its nature and utility may simply seem like a move that obscures rather than clarifies the construct of trust. Again, these criticisms rely on the attempt to apply characterizations and expectations of trust rooted in interpersonal relations to human-AI interactions, raising again questions about the appropriateness of this manoeuvre. We have already discussed this point at the end of Sect. 2. More importantly, the 'trust as a property of reliance' account is not a black-box. In fact, it is possible—and first of all necessary—to provide a characterization of this property based on factors that align with high-level expectations we commonly associated with trust. Such characterization exists and it allows relating trust to the efficiency of the underlying reliance relation it qualifies. Through this characterization we provide a justification of the 'trust as a property of reliance' account and potentially alleviate the aforementioned reservations. We describe it in what follows.

3.3.2 Step II. Trust and Efficiency in Reliance Relations

A characterization of the 'trust as a property of reliance' account can be provided by endorsing two high-level intuitions, that is, intuitions that do not depend on any specific account of trusting, but which are commonly associated with trust.

The first intuition states that, if we trust a lot, we will not exercise a high level of monitoring during reliance. Here, monitoring refers to the allocation of resources aimed at overseeing and potentially intervening in the activities carried out by the

[8] Gambetta also seems to suggest that trust is not a relation but rather a scalar quantity: "trust [...] is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action [...] and in a context in which it affects his own action" (Gambetta, 2000, p. 217). However, in the same work, he later introduces an epistemic account of trust–see Sect. 2–stating that "trusting a person means *believing* that when offered the chance, he or she is not likely to behave in a way that is damaging to us [...]" (Gambetta, 2000, p. 219, emphasis is ours).

trustee.[9] Essentially, if I trust you (a lot), I will not check or intervene on you and your activities (a lot). The idea that trust and monitoring (or control) are somehow antithetic has been discussed over time by various authors (Dasgupta, 1988; Keren, 2014; Baier, 2013).[10]

By adopting the account of trust as a property of reliance, we can formalize this inverse proportionality, capturing in a precise way expressions like 'I trust you a lot/not so much,' which are prevalent in the literature. Taddeo did it in her work on e-trust (Taddeo, 2010), introducing a model that formalizes the inverse proportionality between levels of e-trust and monitoring resources. According to Taddeo, "the second-order-property of e-trust has the effect of minimising an agent's effort and commitment in the achievement of a given goal." (Taddeo, 2010, pag. 244). Then, in Taddeo's account, e-trust encodes the efficiency of reliance in digital contexts.[11]

The second intuition suggests that the degree of trust we place in a trustee is positively correlated with the expected gain we anticipate from the trustee's success. In other words, other things equal, we tend to trust more when there is a substantial potential benefit, net of any possible losses.[12] This intuition has been elaborated in some detail by Loi et al. (2023). In their work, the authors introduce the notion of 'stakes,' defined as the expected gain in a reliance relation, which is a function of the trustor's subjective confidence in the trustee's performance, given a predefined level of monitoring before the interaction begins. The use of expected gains (net of losses) to characterize trusting is reminiscent of the idea that trust occurs in contexts where the trustor is vulnerable, due to uncertainty and risk (Baier, 1986). An increase in the trustor's confidence in the trustee's success leads to an increase in stakes, which, in turn, increases the level of trust in that interaction, other things equal (for instance, the planned investment of resources). See Section 3.5 in Loi et al. (2023) for more details on these dynamics. Then, like monitoring, subjective confidence serves as a measure of the informational efficiency of the reliance relation that unfolds.

To illustrate this point, let us return to the example of Bob and his medical AI. Over time, possibly due to his high levels of monitoring, Bob improves his understanding of the system's functionalities. Although he might reduce the need for monitoring in any instance of appropriate reliance, Bob remains obliged to exercise a certain degree of control over the AI due to the policies enforced by his healthcare institution. These policies mandate a sequence of checks and other control activities that Bob must perform, ultimately leading to the charting of treatment decisions and

[9] Using Castelfranchi and Falcone's words: "[monitoring] aime[s] at ascertaining whether another action has been successfully executed" (Castelfranchi & Falcone, 2010, pag. 193).

[10] Baier is explicit on this point: "As I understand trust, it itself involves economizing on monitoring, supervision, and audits, and leaving the trusted to get on with their work with minimal audits and minimal supervision. So increasing these is of course displaying decreasing trust – simply replacing it with audits, supervision, threats, sanctions and coercion" (Baier, 2013, p. 175). Also Gambetta relates trusting to monitoring—see footnote 9.

[11] However, the monitoring exercised by X in a reliance relation with Y (e.g., an AI system) may provide X with reasons to reduce monitoring in subsequent interactions, thereby relying on Y with progressively higher levels of trust over time.

[12] We may err in estimating these benefits, or place high levels of trust in an entity without good reason—for instance, when that entity has little or no capacity to perform the action we expect of it. More on this in Sect. 3.4.

their rationale. Does this mean that Bob would never be able to rely on the AI appropriately with high levels of trust? Not necessarily. Even with the mandated monitoring requirements, Bob's trust in the medical AI can increase if his confidence in the system grows because of efficiently managing the information he acquires over time. In summary, high levels of trust in a reliance relation can result from two dynamics, either an optimization of monitoring resources, or an increase in the subjective confidence in the AI system's ability to generate accurate predictions. Loi and colleagues extended Taddeo's model by introducing two axioms that formalize this relationship between trust, monitoring, and stakes, enabling the derivation of mathematical functions to describe levels of trust within a given reliance relation—see Sects. 3.5 and 5 in (Loi et al., 2023) for all details.

A final remark to distinguish our approach from those of Taddeo and Loi et al. and to conclude this section. On the one hand, Taddeo introduces trust as a property of relations within the context of e-trust, framing it as the result of a rational choice made by an artificial agent communicating with other artificial agents (Taddeo, 2010). Loi et al. extend Taddeo's approach by (1) defining trust as a quantitative property of *all* reliance relations and (2) considering trust as antithetical to monitoring and proportional to stakes. Their goal is to model levels of trust (Loi et al., 2023). On the other hand, while our approach is clearly inspired by these, it diverges from them by arguing that *the formalization of trust as a property of reliance and its quantification are justified by the pragmatics of reliance*. Specifically, this perspective enables us to account for the requirement of efficiency, as discussed in Sects. 3.3.1 and 3.3.2. In fact, efficiency, defined as the ability to achieve desired outcomes with minimal waste of time, effort, or resources, is attained when one appropriately relies on an AI system with high *levels* of trust. Finally, our approach allows us to measure levels of trust in relation to the objective conditions that make an AI system worthy of trust, as we will explore in the forthcoming section.

3.4 Relying on an AI for the Right Reasons: Calibrated Reliance

Striving for appropriate and efficient reliance relations with AI is not enough. Especially in high-risk contexts—see the recent EU AI Act for a classification of these (European Commission, 2024), it is key to ensure that the accuracy of the AI-assisted human decision-maker and the user's high levels of trust in the system are rooted in the right reasons. Intuitively, as we are dealing with machines, these reasons should be tied to the objective capabilities of the AI system to perform as expected, starting with keeping an accurate predictive power. The idea that relying on AI results from the system's actual capabilities can be referred to as 'calibration.'

Interestingly, the literature on human-AI interactions reserves this term exclusively to the expression 'trust calibrated in trustworthiness.' This expression has been introduced by Lee and See in their work on reliance and trust in automation (Lee & See, 2004), where the authors define calibration as "the correspondence between a person's trust in the automation and the automation's capabilities" (Lee & See, 2004, pag. 55).[13] If we refrain from defining 'trust' and characterizing the 'correspondence'

[13] See also the references therein.

for a moment, here, trustworthiness is understood as the automation's ability to perform reliably and as expected—its objective capability of being trusted.

The idea that the trustworthiness of AI includes objective capabilities of these systems is widespread; it aligns with the notion that being worthy of trust in interpersonal relations involves demonstrating competence and reliability among others (Baier, 1986; Dasgupta, 1988; Pettit, 1995). The question is whether objective capabilities do suffice to define trustworthiness or other factors need to be considered as well. Critics who argue that AI cannot be genuinely trusted often dismiss trustworthiness as an empty concept, given that AI lacks the capabilities, for instance, mental states, required for genuine trust. For them trustworthiness of AI is a glorified form of reliability. In contrast, scholars endorsing the 'Trustworthy AI' framework promote the idea that the objective capabilities constituting the trustworthiness of AI descend from principles that govern a social system (Floridi et al., 2018). For instance, the European Commission's High-Level Expert Group on Trustworthy AI[14] focuses on key principles, such as respecting human agency, ensuring technical robustness and safety, and promoting transparency. that typically include accuracy, fairness, robustness, reproducibility, and explainability among others (Kaur et al., 2022). As a result, the trustworthiness of an AI system is a composite, context-dependent property of the system (Mattioli et al., 2024; Ferrario, 2024). The properties, or capabilities, that constitute this trustworthiness are defined by the system designers. Users assess the trustworthiness of an AI system through direct evaluations, audits, accessing third-party certifications, and social transparency cues instead (Ehsan et al., 2021). To the 'Trustworthy AI' framework belongs the 'contractual trust' approach to the trustworthiness of AI, which promotes the idea that an AI system is worthy of our trust if and only if it works as it is supposed to in its context of application (Jacovi et al., 2021). This approach descends from the work of Hawley's 'trust with commitments' idea, namely, that to trust is to believe that the AI will maintain the ascribed functionalities, or 'contracts' (Hawley, 2014). In human-AI interactions, the capabilities to be maintained constitute the trustworthiness of the AI, which is defined and translated into technical requirements by its designers depending on the context of its use. Thus, from an engineering perspective, trustworthiness is a second-order, principle-inspired property of the system that comprises all the functionalities that the AI is supposed to maintain (to a certain level) during its use in a given context.

Back to our pragmatic account and 'trust as a property of reliance.' The reader has surely noticed that the notion of 'calibrating trust in trustworthiness' is conceptually problematic: we aim to develop AI systems that are worthy of our trust, anchoring trust in their virtuous capabilities, but we still lack a coherent account of what 'trust in AI' actually means. Without solving this circularity, 'calibrating trust in trustworthiness' remains a proposition possibly providing useful recommendations to AI designers, but without an appropriate theoretical framework. Our pragmatic approach and formalization of trust as a property of reliance offers a solution to this impasse. What we actually seek is for our appropriate reliance to stem from accurate perceptions of an AI system's objective capabilities, met at a

[14] Their guidelines are available here: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

level judged adequate. In other words, the real task here is to *calibrate reliance on trustworthiness*, where the calibration implies that highly trustworthy AI systems induce high-levels of trust. We define the trustworthiness of AI as the capability of an AI system that gives reasons to rely on the system with high levels of trust. Equivalently, a trustworthy AI is one that possesses a capability providing reasons to rely on the system while efficiently allocating resources–see Sect. 3.3.2. Our view is consistent with the idea that trust is antithetic to monitoring and correlates with the human's confidence in the capability of the AI system to generate accurate predictions (Loi et al., 2023). As in our AI-assisted decision-making approach, to rely is equivalent to make the decision to use the AI's prediction as opposed to relying on one's own, then, the capability grounding trustworthiness is constituted by the functionalities that promote accuracy of the AI. Beyond optimizing predictive accuracy, these comprise (the technical implementations of) safety, reliability, robustness, fairness, and transparency among others. Thus, a trustworthy AI is reliable, though reliability alone, however, does not equate to trustworthiness and reliance is calibrated in trustworthiness when to rely on an AI's outputs is caused by its trustworthiness.[15] For instance, Bob would have not relied on the medical AI for a difficult to classify medical image if he had not accessed the ad-hoc interface with the technical requirements that allowed him to form a mental representation of the property of being accurate. The technical requirements support the formation of the mental representation as they provide information on the predictive accuracy, safety, and reliability levels of the system. Accessing the interface requires Bob to invest monitoring resources, which reduces his level of trust in the AI system compared to a scenario without such access. However, the mental representation Bob acquires through this reliance relation may later justify relying on the system with reduced or no monitoring, thereby increasing his subsequent levels of trust in it.[16]

4 Implications of our Pragmatic Approach

Our pragmatic approach to trust in AI grounds the ontological repositioning of trust from a relation to a property of another relation, i.e., reliance, not in the rationality of interactions between artificial agents (Taddeo, 2010), nor as an attempt to model quantitative levels of trust (Loi et al., 2023), but rather within the pragmatic expectations of reliance. As a result, it allows us conceptualizing trust in AI as genuinely distinct from 'mere reliance' on these systems while

[15] More precisely, we should say 'caused by an appropriate assessment of the objective capabilities of the AI.' In other words, we have to align the perception of trustworthiness with the (objective) trustworthiness of the system, providing justification of the belief on the trustworthiness of the system that results. As the perception of the trustworthiness of the system is caused by the interaction with the AI, process reliabilism seems a good starting point (Ferrario, 2024).

[16] This said, calibrated reliance is not necessarily accurate. This is because modern AI systems use probabilistic methods to generate predictions. Thus, trustworthiness is held up to a certain standard, which may degrade over time. As a result, despite a correct assessment of the objective trustworthiness of the system, the possibility of idiosyncratic errors may lead to inaccurate instances of reliance.

enabling the formalization of how reliance is calibrated in the trustworthiness of AI systems. As a result, calibrating reliance in the trustworthiness of AI becomes a theoretically-sound and empirically measurable procedure that involves manipulating the properties that constitute the system's trustworthiness and measuring variations in reliance.[17] This is an important practical implication of our approach. It allows researchers to identify which objective capabilities of an AI system influence user reliance in empirical studies. In fact, focusing on calibrating appropriate reliance in trustworthiness presents a more straightforward and less contentious path than calibrating trust, which would require reaching consensus around the nature of trust and its quantification. The ongoing philosophical debate shows that such consensus is yet to be reached, while systematic reviews of the empirical research on trust in AI offer a rather discouraging picture of the multitude of methods, including one-item, ad-hoc questionnaires, and behavioural assessments—all used to measure trust in this technology (Benk et al., 2024). By calibrating appropriate reliance in the trustworthiness of AI we rely on AI technology appropriately as the consequence of a correct assessment of the objective capabilities that make this technology worthy of being relied upon. In summary, our shift toward a reliance-centered perspective in a traditionally trust-dominated discourse offers a pragmatic and empirically sound approach for investigating how people interact with AI technology. A short example. Let us suppose researchers want to examine which AI features drive the recalibration of reliance by Bob (and his colleagues) over time. To do so, they want to test whether specific trustworthiness properties, such as explainability modules or fairness constraints, foster trust, in the sense of increasing reliance by reducing the monitoring effort Bob and his colleagues consider necessary over time. Such a study could, for instance, compare a standard AI system with one augmented by an explainability module that generates counterfactual explanations through an LLM-based natural language interface. This design would make it possible to assess whether the intervention increases instances of appropriate reliance and reported trust levels by Bob and his colleagues across repeated use, even if it initially decreases trust due to the additional resources invested by them to explore the explanations. Applied systematically, this methodology would enable AI providers and users to document how design choices shape reliance trajectories, to detect, over time, when reliance deviates from appropriate calibration in a given context of use and implement mitigation strategies.

5 Conclusions

In this paper, we have argued for a pragmatic approach to defining trust in AI that is grounded in three fundamental expectations that we believe should guide human-AI interactions: appropriate reliance, its efficiency, and calibration. Our reliance-centered approach draws a formal distinction between reliance and trust: it does not reject the concept of trust, but repositions it as a a quantitative property of reliance.

[17] See (et al., 2023) for a measure of appropriate reliance.

We have shown that by focusing on these expectations, it is possible to reconcile reliance with trust in a way that is both theoretically sound and practically useful. In particular, our approach grounds the trust-as-antimonitoring model attempts by (Taddeo, 2010) and (Loi et al., 2023) in pragmatic expectations of reliance. By re-framing trust as a quantitative property of reliance, we have sought to move beyond the traditional philosophical dichotomies that have long characterized the debate. This approach offers a viable alternative to rational or motivational accounts of trust in AI, as well as to arguments denying the possibility of trust in AI. It offers a philosophical perspective on trust in AI that is both actionable and aligned with empirical research. That said, some limitations remain. We believe there will always be scholars who defend a characterization of trust in AI modeled on interpersonal trust, emphasizing mental states, moral obligations, affective commitments, or virtues, especially given the persuasive and anthropomorphic affordances of large language model-based applications (Floridi & Nobre, 2024; Shanahan et al., 2023). Others may not be persuaded by our ontological repositioning of trust from a relation to a property of a relation, even if we grounded this move in the expectations of reliance, as this approach shifts the traditional normative emphasis of trust towards the calibration of reliance in AI trustworthiness. Finally, although trust as a property of reliance can be modeled, for instance, following the approach of Loi et al. (2023), important elements still require further conceptual development, such as the account of subjective probability formation underlying the stakes estimate, while systematic empirical validation of this quantitative framework to trust has yet to be carried out. As AI becomes more integrated into decision-making processes, particularly in high-stakes environments like healthcare, being pragmatic about reliance and trust in AI should not be dismissed as merely technical or philosophically shallow. Instead, it offers a practical and meaningful framework for addressing the nuances of trust in AI.

Acknowledgments The author thanks the anonymous reviewers for their constructive feedback and Michele Loi and Alessandro Facchini for insightful conversations.

Author Contributions Not applicable.

Funding Open access funding provided by University of Zurich. None to declare.

Data Availability Not applicable.

Declarations

Conflict of interest None to declare.

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

Alvarado, R. (2023). What kind of trust does AI deserve, if any? *AI and Ethics, 3*(4), 1169–1183.

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

Baier, A. 2013). What is trust? In D. Weinstock, N. Manson, M. Deveaux, & D. Archard (Eds.), *Reading Onora O'Neill* (pp. 175–185). Routledge.

Benk, M., Kerstan, S., von Wangenheim, F., & Ferrario, A. (2024). Twenty-four years of empirical research on trust in AI: A bibliometric review of trends, overlooked issues, and future directions. *AI & Society, 40*(4), 2083–2106.

Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philosophy & Technology, 34*, 349–371.

Castelfranchi, C., & Falcone, R. (2010). *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley & Sons.

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology, 14*, 53–60.

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Cogley, Z. (2012). Trust and the trickster problem. *Analytic Philosophy, 53*(1), 30–47.

Dasgupta, P. (1988). Trust as a commodity. D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 49–72). Blackwell.

DeCamp, M., & Tilburt, J. C. (2019). Why we cannot trust artificial intelligence in medicine. *The Lancet Digital Health, 1*(8), Article e390.

Durán, J. M., & Pozzi, G. (2025). Trust and trustworthiness in AI. *Philosophy & Technology, 38*(1), 16.

Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O., & Weisz, J.D. (2021). Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–19).

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.

European Commission (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act).* (Regulation 2024/1689; Official Journal of the European Union, L, 12 July 2024)

Fabris, A. (2020). Can we trust machines? The role of trust in technological environments. *Trust: A Philosophical Approach* (pp. 123–135). Cham, Switzerland: Springer.

Ferrario, A. (2024). Justifying our credences in the trustworthiness of AI systems: A reliabilistic approach. *Science and Engineering Ethics, 30*, 55.

Ferrario, A., Facchini, A., & Termine, A. (2024). Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds and Machines, 34*(3), 30–27.

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology, 31*(1), 1–8.

Floridi, L. (2020). *The ethics of AI and robotics: A philosophical introduction*. Oxford University Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4people—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707.

Floridi, L., & Nobre, A. C. (2024). Anthropomorphising machines and computerising minds: The crosswiring of languages between artificial intelligence and brain & cognitive sciences. *Minds and Machines, 34*(1), 5.

Gambetta, D. (2000). Can we trust trust? D. Gambetta (Ed.), *Trust: Making and Breaking Cooperative Relations* (pp. 213–237). Oxford: Basil Blackwell.

Goldberg, S.C. (2020). Trust and reliance 1. *The Routledge Handbook of Trust and Philosophy*, 97–108.

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2011). Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". *Ethics and Information Technology, 13*, 17–27.
Hardin, R. (1992). The street-level epistemology of trust. *Analyse & Kritik, 14*(2), 152–176.
Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation.
Hatherley, J. J. (2020). Limits of trust in medical AI. *Journal of Medical Ethics, 46*(7), 478–481.
Hawley, K. (2014). Trust, distrust and commitment. *Noûs, 48*(1), 1–20.
Hawley, K. J. (2019). *How to be Trustworthy*. Oxford University Press.
Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy, 72*(1), 63–76.
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* (pp. 624–635).
Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR), 55*(2), 1–38.
Keren, A. (2014). Trust and belief: A preemptive reasons account. *Synthese, 191*(12), 2593–2615.
Lai, V., Chen, C., Liao, Q.V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. arXiv preprint arXiv:2112.11471
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
Loi, M., Ferrario, A., & Viganò, E. (2023). How much do you trust me? A logico-mathematical analysis of the concept of the intensity of trust. *Synthese, 201*(6), 186.
Mattioli, J., Sohier, H., Delaborde, A., Amokrane-Ferka, K., Awadid, A., Chihani, Z., & Pedroza, G. (2024). An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering. *AI and Ethics, 4*(1), 15–25.
Mollering, G. (2006). *Trust*. Emerald Group Publishing Limited.
Nickel, P. J. (2007). Trust and obligation-ascription. *Ethic. Theory Moral Prac., 10*(3), 309–319.
Peirce, C.S. (1940). How to make our ideas clear. J. Buchler (Ed.), *Philosophical Writings of Peirce* (pp. 23–41). New York: Dover Publications.
Pettit, P. (1995). The cunning of trust. *Philosophy & Public Affairs, 24*(3), 202–225.
Pettit, P. (2004). Trust, reliance and the internet. *Analyse & Kritik, 26*(1), 108–121.
Primiero, G. (2020). A logic of negative trust. *Journal of Applied Non-Classical Logics, 30*(3), 193–222.
Primiero, G., & Taddeo, M. (2012). A modal type theory for formalizing trusted communications. *Journal of Applied Logic, 10*(1), 92–114.
Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics, 26*(5), 2749–2767.
Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 410–422).
Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*(7987), 493–498.
Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., & Shen, D. (2020). Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Reviews in Biomedical Engineering, 14*, 4–15.
Simon, J. (2020). *The Routledge Handbook of Trust and Philosophy*. Routledge.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
Taddeo, M. (2009). Defining trust and e-trust: From old theories to new problems. *International Journal of Technology and Human Interaction (IJTHI), 5*(2), 23–35.
Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*, 243–257.
Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.
Van Wynsberghe, A., & Robbins, S. (2014). Ethicist as designer: A pragmatic approach to ethics in the lab. *Science and Engineering Ethics, 20*(4), 947–961.
Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–39.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D.L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189–201).

Zanotti, G., Petrolo, M., Chiffi, D., & Schiaffonati, V. (2023). Keep trusting! A plea for the notion of trustworthy AI. *AI & Society*, 1–12.