AI & SOCIETY
https://doi.org/10.1007/s00146-023-01789-9

MAIN PAPER

Keep trusting! A plea for the notion of Trustworthy AI

Giacomo Zanotti[1] · Mattia Petrolo[2,3] · Daniele Chiffi[4] · Viola Schiaffonati[1]

Received: 24 April 2023 / Accepted: 18 September 2023
© The Author(s) 2023

Abstract

A lot of attention has recently been devoted to the notion of Trustworthy AI (TAI). However, the very applicability of the notions of trust and trustworthiness to AI systems has been called into question. A purely epistemic account of trust can hardly ground the distinction between trustworthy and merely reliable AI, while it has been argued that insisting on the importance of the trustee's motivations and goodwill makes the notion of TAI a categorical error. After providing an overview of the debate, we contend that the prevailing views on trust and AI fail to account for the ethically relevant and value-laden aspects of the design and use of AI systems, and we propose an understanding of the notion of TAI that explicitly aims at capturing these aspects. The problems involved in applying trust and trustworthiness to AI systems are overcome by keeping apart trust in AI systems and interpersonal trust. These notions share a conceptual core but should be treated as distinct ones.

Keywords Trustworthy AI · Ethics of AI · Trust · Reliance · Interpersonal and artificial trust

1 Introduction

In the last decades, interpersonal trust and trustworthiness have often been the subject of philosophical debate (see McLeod 2021; Simon 2020). In addition, the notions of trust and trustworthiness have also been employed to characterize the nature of the relationship between humans and AI systems (for an overview, see Grodzinsky et al. 2020; Taddeo 2009). Importantly, they play a pivotal role in the context of the European Union's ethics-based effort to provide a regulatory framework for the design and use of AI systems. Most notably, the much-discussed European proposal for the *Artificial Intelligence Act* (European Commission 2021) is preceded and inspired by the *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019).

That being said, a consensus on what makes an AI system trustworthy is still missing. And understandably so, one might add. After all, discussions on trust and artificial systems are fairly recent, and technological development in the field of AI is rapid. Most importantly, however, the development of the debate is hampered by a foundational problem. While it is relatively uncontroversial that humans can be trustworthy, the same cannot be said for AI systems. More precisely, it has been argued that the concept of trust is unfit for describing the interaction between humans and artificial agents, and that the very notion of Trustworthy AI – hereinafter, TAI – is a categorical mistake. These concerns have been voiced, among others, by Thomas Metzinger, a member of the expert group set up by the European Commission that worked on the *Ethics Guidelines*. In Metzinger's (2019) view, the "underlying guiding idea of a 'trustworthy AI' is, first and foremost, conceptual nonsense", for "only humans can be trustworthy (or untrustworthy)". At this stage, a certain discomfort seems justified. Although not flawless, the European regulatory effort on AI has generally been praised as a crucial step towards a safe and responsible deployment of AI systems. And yet, the conceptual and ethical foundation of this effort might be irremediably compromised, for the *Ethics Guidelines* at the basis of the European legislative proposal on AI explicitly build upon the notion of TAI.

✉ Giacomo Zanotti
giacomo.zanotti@polimi.it

✉ Viola Schiaffonati
viola.schiaffonati@polimi.it

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

[2] Centre for Philosophy of Science of the University of Lisbon (CFCUL), Campo Grande, Building C4 1749-016, Lisbon, Portugal

[3] Federal University of ABC, Alameda da Universidade, s/n 09606-045, São Bernardo Do Campo, SP, Brazil

[4] Department of Architecture and Urban Studies, Politecnico di Milano, Milan, Italy

Published online: 12 October 2023

The aim of this paper is twofold. First, we provide an overview of the debate on the notion of TAI, relating it to the general philosophical debate on trust and shedding light on its desiderata and implicit assumptions. Second, we argue that the main views on TAI prevent us from making good sense–and use–of this notion. We propose to use the notion of TAI for capturing the ethically relevant and value-laden aspects of the design and use of AI systems, and we provide a conceptual framework that allows to meaningfully and legitimately apply the notion of trust to AI systems.

Before starting, a couple of caveats are in order. First of all, we need to be aware of semantic overlaps and avoid linguistic confusion. As we will see in the next section, the debate we are considering largely revolves around the contraposition between two notions: trust and reliance. Without anticipating too much, the former is used in the context of interpersonal interactions, whereas the second is typically used to describe our relationship with inanimate objects and tools. The whole point, here, is whether AI systems represent an exception and can be trustworthy. In everyday language, however, trust and reliance are pretty much interchangeable. For example, it makes sense to say that I trust my scale, or that I rely on my cardiologist. Among other things, this makes it difficult to test our intuitions concerning trust and AI systems.

The second aspect that should not be overlooked has to do with the risk of idealization. Much of the literature on general trust aims at specifying its determinants. Depending on the kind of account in question, these determinants can be purely epistemic in nature or involve some normative and motivational dimensions. This approach, which conceives trust as the result of some clearly identified factors, is largely inherited by the debate on TAI. However, we should never forget that, in real-world contexts, trust is a complex psychological phenomenon that might not be exhaustively captured by philosophical analysis. Note that this is not to say that philosophical work on trust is flawed or useless. More modestly, it should be kept in mind that there might be a difference between ideal and real-world trust, for the second can be subject to the influence of contingent factors – for instance, there might be intercultural variations (see Klein et al. 2019). So, while philosophical analysis might provide us with an understanding of the distinctive features of trust, some room should be left for contextual factors. Aware that real-world trust might not be exhaustively captured by a fixed set of determinants, we can turn to the philosophical debate on trust and trustworthy AI.

In Sect. 2, we provide an overview of the philosophical debate on trust, focusing on the contraposition between purely epistemic and motivational accounts of trust. After that, in Sect. 3, we present the main views on trust in AI originating from these accounts. In Sect. 4, we argue that these views prevent us from using the notion of TAI in a meaningful way, which clearly has a negative impact on the attempt to provide a regulation for AI that is ethically based on trust. We propose an alternative view that insists on the importance of grounding the notion of TAI in the ethically relevant and value-laden aspects of the design and use of AI systems. Finally, in Sect. 5, we specify the relation between interpersonal trust and trust in AI systems, arguing that they should be treated as two distinct—although related—notions. Section 6 contains a brief summary of the paper and some concluding remarks.

2 Understanding trust

Let us start by specifying the structural features of trust and trustworthiness. To begin with, trust is a relational matter between the trusting party, or trustor, and the trusted party, or trustee. I can trust someone—let us temporarily focus on interpersonal trust—but it makes little sense to say that I trust *simpliciter*.[1] On the contrary, trustworthiness is a non-relational property. Indeed, it makes perfect sense to say that someone is trustworthy without any further reference to possible trustors. That being said, the link between the two notions is usually taken to be trivial: being trustworthy roughly means deserving trust. Accordingly, trust and trustworthiness have systematically been addressed together in the literature.[2]

Although the general structure of trust and trustworthiness is relatively uncontroversial, filling in the details is more complicated. In particular, authors disagree on the nature of trust and therefore on what it takes to be trustworthy. A shared assumption is that trust requires reliance and, accordingly, trustworthiness requires reliability. Instances of reliance are widespread in our ordinary life: we rely on seat belts to protect us in the unfortunate circumstance of an accident, we rely on heating systems to keep us warm during winter, and so on. A clear definition of reliance is provided by Goldberg (2020, p. 97): "where 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". Though necessary, mere reliance does not seem to be enough for trust. In general, most authors agree that reliance is the basis of trust. However, they also agree on the fact that some extra element is required in addition to

[1] At most one can be *trusting*, but that simply refers to one's tendency to trust people.

[2] Note that Buechner et al. (2014) argue that the relation between trust and trustworthiness is underexplored. For the purpose of this paper, however, we maintain that trustworthiness is the property x must possess in order to be justifiably trusted.

reliance. When it comes to specifying the nature of this extra element, disagreement begins.

An overview of the philosophical literature on general trust goes way beyond the scope of this paper, for our discussion specifically aims at targeting the notion of TAI. Our first aim, let us recall it, is to assess whether the notion of trust is applicable to AI systems. In this perspective, it will be sufficient to note that the debate on trust largely builds upon the contraposition between two main families of views. For the purpose of this paper, let us call them *epistemic* and *motivational* accounts of trust.[3]

According to epistemic accounts, trusting is a matter of rational choice (e.g., Gambetta 1988; Mollering 2006). The details vary depending on the specific account that is considered. What remains unchanged is that, from an epistemic perspective, trusting X to do φ requires operations such as estimating the probability that X will do φ and evaluating the pros and cons of relying on X to do φ.

The main problem with epistemic accounts is that they have trouble distinguishing between trust and reliance. On the one hand, when I choose to *rely* on something – my laptop's battery, for example – a number of rational considerations are available. I can make inductive considerations on the functioning of the battery, or (roughly) estimate the likelihood that it will work by taking into account its last charging cycle, its conditions, its age, and so on. On the other hand, if trust is just a matter of rational deliberation to act upon the presupposition that someone or something will behave as they are expected to do, what differentiates it from reliance?

A better ground for the distinction between trust and reliance is provided by motivational accounts. The label is admittedly imprecise and captures a number of different theoretical options. The common feature of these accounts is that they take trust to be the combination of reliance and some other extra factor that is not purely epistemic in nature. According to Baier's (1986) influential account, for instance, an essential component of trust is that we choose to rely on someone *under the assumption of their goodwill*. When trusting X to perform a certain task φ, we do not merely act on the presupposition that X will do φ, but rather on the presupposition that X will do φ out of goodwill—or at least, that X is not willing to harm us by purposely not doing φ. Other accounts focus on the trustee's *interests* (Hardin 2002) and *moral obligations* (Nickel 2007).

Note that the epistemic component of rational assessment is not absent in motivational accounts. First of all, it plays a role in the ascription of capabilities to the trustee. We may well assume that X is animated by goodwill and still not trust them to do φ because we lack good reasons to believe that they *are able* to do φ. What is more, rational considerations are needed to evaluate the trustee's motivations and adherence to moral obligations. The difference between epistemic and motivational accounts of trust rather concerns the *object* of rational assessment. As a matter of fact, what undergoes rational scrutiny in epistemic accounts is just the likelihood that X will do φ, as well as X's competence to do φ.

3 Trust in AI

The taxonomy just provided is fairly schematic and may fail to capture differences among accounts. However, it provides a sufficient background for addressing the question of trust in AI, for the disagreement on the notion of TAI is largely dependent on the choice between motivational and epistemic accounts of TAI. Let us consider again the criticism of the very idea of TAI according to which only humans can be trustworthy and therefore the notion of TAI is "conceptual nonsense". However inflammatory, this criticism captures a crucial point. That is, under some prominent understandings of trust, the idea of trust in AI systems and therefore the notion of TAI is highly problematic.

Unsurprisingly, difficulties arise when motivational accounts of trust are accepted, and trust is understood as involving the trustee's motivations, goodwill, and adherence to moral obligations. Once this view is accepted, an argument to the effect that AI systems cannot be trustworthy is easily built:

1. An entity X is trustworthy only if X has the right motivations, goodwill and/or adheres to moral obligations towards the trustor;
2. AI systems lack motivations, goodwill, and moral obligations;
3. Therefore, AI systems cannot be trustworthy.

Arguments of this sort are quite common in the literature on trust and AI (DeCamp and Tilburt 2019; Fossa 2019; Hatherley 2020; Ryan 2020). Sure, there are variations due to the fact that different authors have slightly different conceptions of trust. However, the argumentative strategy is more or less the same: trustworthiness depends on the possession of paradigmatically human features that AI systems lack, and therefore the notion of TAI does not make sense from a conceptual point of view. Given their roots in motivational conceptions of trust, let us call these views *motivational accounts of TAI*. The objection they raise against the

[3] Note that the taxonomies and the labels are not fixed, even if we narrow it down to the literature on trust and AI. For instance, epistemic accounts are called "rational" by Nickel et al. (2010) and Ryan (2020). What we call motivational accounts, instead, are further distinguished into will-based and affective accounts by Beuchner et al. (2014), and into affective and normative accounts by Ryan (2020).

notion of TAI, instead, will be referred to as the *conceptual nonsense* objection. Importantly, the conceptual nonsense objection can also be motivated by concerns about the undue anthropomorphization of artificial agents, for the discourse on TAI would suggest that AI systems actually possess the paradigmatically human features that are typically associated with trustworthiness.[4]

Now, two points are worth making explicit. First, the central claim of motivational accounts of TAI is not that AI systems are untrustworthy. A person animated by ill will or interested in hurting us is untrustworthy, but it is beyond doubt that the *question* on their trustworthiness makes perfect sense, at least from a conceptual point of view. What motivational accounts of TAI reject is the very applicability of considerations of trustworthiness to AI systems. Since these systems lack relevant features such as motivations, will, and moral obligations, saying that they are trustworthy or untrustworthy would be a categorical mistake. It would be like ascribing colours to numbers or emotions to tables.

Second, one could object that *currently available* AI systems lack these features, but there is no principled reason why they could not possess them in the future. For instance, one could think about a scenario in which some AI systems become full ethical agents in Moor's (2006) sense: agents with intentions, consciousness, and free will, who are able to make and justify fully-fledged ethical judgments. These systems would arguably qualify as trustworthy – or untrustworthy, if one is persuaded by catastrophist narratives on an AI takeover. The possibility of AI systems attaining the level of full ethical agents is debated in machine ethics (Anderson and Anderson 2011; Hunyadi 2019). Now, this question falls outside the scope of this paper, and we will not address it here. What we wish to stress is instead the fact that, at best, full ethical artificial agents are nothing but a conjecture, at least at the moment. In contrast, the debate on TAI—as well as the European regulatory effort that is based on this notion (see Mökander et al. 2022)—is all about current AI systems. We stick to this latter approach, and in what follows we will just assume that AI systems do not possess motivations, will, and moral obligations. Importantly, this does not prevent us from considering ethical questions concerning AI. On the contrary, in the next section, we will insist on ethically relevant and value-laden aspects of the design and use of AI systems.

That being said, motivational accounts of TAI are not the only theoretical option. In fact, even if the possibility of full ethical artificial agents is dismissed, some have argued that AI systems can be trustworthy. Although not always made explicit, the theoretical ground for this second family of views are epistemic accounts of trust. Accordingly, we call them *epistemic accounts of TAI*. Again, the details vary. In general, however, we can say that epistemic accounts of TAI apply trust and trustworthiness to AI systems on the basis of these systems' performance and the justification of our beliefs about it.

To put some flesh on the bones, consider the view advocated by Ferrario et al. (2020, 2021), focusing on their discussion of medical AI systems. Trust in these systems is defined as the "reliance property that describes the willingness of the physician to rely on the medical AI without intentionally generating and/or processing further information about the medical AI's capabilities to achieve the goal at hand (e.g., by monitoring the medical AI)" (2021, p. 437).[5] According to this view, motivations and moral obligations play no role in the building of trust. On the contrary, trust in AI is conceived as a purely epistemic phenomenon that involves two steps. First, there is the "mere reliance" phase, in which we rely upon a certain artificial system and come to entertain some beliefs about its performance. These beliefs are constantly updated through repeated interactions with the system—as well as exchanges with experts and other users. At some point, we stop updating these beliefs and just rely on the system. When this happens, the system is trusted.[6] Basically, the difference between mere reliance and trust would be the presence of the monitoring activity.

Again, this is not the only way to understand trust in AI in epistemic terms. Among others, Durán and Jongsma (2021) and Durán and Formanek (2018) advocate a form of

[4] The issue of anthropomorphism in AI is indeed an urgent one, especially when AI systems are used in contexts where they end up playing roles that are typically played by humans – for instance, we can think about robots in health care or AI-powered chatbots that are designed and marketed as "virtual friends", such as Replika (Skjuve et al. 2021). In general, the way we talk about AI systems heavily contributes to the process of anthropomorphization. On the one hand, referring to these systems' functioning in terms of paradigmatically human abilities (understanding, reasoning, and so forth) may relieve us from explaining the complex details of their working principles. On the other hand, this practice might encourage the erroneous ascription of human-like capabilities to AI systems. In principle, something analogous might happen with trust, for it is assumed that trust can only occur in interpersonal relations. Now, anthropomorphism is arguably a largely psychological problem. Yet, we believe it is one we should take into account when providing specific ethical recommendations for the design and use of AI systems. Providing such recommendations falls outside the aims of this work, which rather seeks to outline a conceptual framework for the applicability of the notions of trust and trustworthiness to AI systems. However, our analysis should lay the conceptual groundwork for more specific recommendations, desiderata and requirements that should contribute to making AI systems "trustworthy", including those tackling the risk of anthropomorphization.

[5] For a similar view, see Nguyen (2022).

[6] In this account, this qualifies as *simple* trust. When simple trust is combined with the belief that simple trust is justified or appropriate, then we have *reflective* trust.

reliabilism, dubbed “computational reliabilism”, that specifies the conditions of epistemic reliability under which an algorithm’s output is trustworthy.[7] What remains unvaried is that, contrary to motivational accounts of TAI, epistemic accounts of TAI make the notion of trustworthiness fully applicable to AI systems by grounding it in a purely epistemic dimension.

For the sake of completeness, let us note that the literature on TAI is very large and constantly expanding, and conceptualizations of trust in its application to AI abound. Some of these are different from the views mentioned so far. For example, Starke et al. (2022) maintain that some kind of intentionality is needed for something to be trustworthy, but weaken the concept of “intention” in a way that makes artificial systems potentially trustworthy. (Coeckelbergh 2012), instead, focuses on “virtual trust” in robots, based on robots’ *appearance* of trustworthiness. Such accounts would deserve a separate discussion, for they hardly fit into the taxonomy of views we are considering.

In the rest of this paper, we will narrow it down to motivational and purely epistemic accounts of TAI. Besides largely motivating the debate we are considering, the contraposition between these two views perfectly captures the core problems raised by the notion of TAI. As a matter of fact, two points have so far emerged in the discussion. On the one hand, unsurprisingly, we want to make sense of the notion of TAI. In this perspective, a minimum requirement is that it is not conceptual nonsense. On the other hand, in line with the literature on general trust, there is widespread agreement that TAI should be different from merely reliable AI. The contraposition between motivational and purely epistemic accounts of TAI largely reduces to the tension between these two desiderata.

4 Trust beyond algorithmic performance

In this section, we will argue that both epistemic and motivational accounts of TAI show some limitations in that they end up preventing the notion of TAI from either being usable or capturing crucial value-laden aspects of the design and use of AI systems. These latter aspects will be the starting point for our positive proposal.

As we have seen, purely epistemic views have the merit of allowing for an unproblematic notion of TAI, at least to the extent that it does not turn out to be conceptual nonsense. Since they do not ground trustworthiness in features that AI systems do not possess, such as motivations, goodwill, and moral obligations, the resulting notion of TAI is not conceptually flawed or at best unrealistic vis-à-vis the current status of research in AI and its foreseeable development. On the contrary, views that build on motivational accounts of trust stand against the very notion of TAI, which is taken to be a categorical error.

The situation is reversed when it comes to the distinction between trust and reliance. As pointed out above, motivational accounts conceive of trust in terms of reliance *plus* some other factor. Importantly, this other factor is not purely epistemic in nature. On these grounds, the distinction between trust and reliance is easily guaranteed. Things are more controversial when it comes to purely epistemic accounts of TAI. Consider again the view according to which trust in AI begins when we stop updating our beliefs about the performance and error patterns of the system in question, and the reliance on the system becomes unmonitored. Having in mind motivational accounts, one might wonder what differentiates trust from reliance in this view. Its advocates’ reply is not hard to guess: the fact that it is unmonitored. The following question comes spontaneously: is this enough to ground the distinction?

The whole discussion quickly turns into a clash of intuitions and preliminary assumptions, for the answer to this question largely depends on what account of general trust is more or less explicitly presupposed. However, not only is there no agreement on the intension—that is, the precise meaning—of the notion of trust. There also seems to be disagreement on its extension—that is, the set of things it applies to. As a matter of fact, without a preliminary understanding of trust, it is hard to distinguish between trust-based and not trust-based (e.g., merely reliance-based) relationships. In other words, there is no agreement on the kinds of relations that should be subsumed under the notion of trust, and therefore we cannot decide between different accounts of trust by checking whether they capture all and only the right, so to say, relations. Needless to say, these disagreements about trust have immediate implications when it comes to discussions on TAI.

In light of these considerations, it is quite dubious that the debate on the notion of TAI can be settled just by contraposing different accounts. This stalemate, however, is not only undesirable from a theoretical perspective. It also has a negative impact on a practical matter of the utmost importance, namely the ethics-based and trust-based attempt to provide a regulation for AI. Here, we advocate the necessity of a different approach. Instead of starting with our preferred account of general trust and then trying to apply it to AI systems, we begin with a question: what advantages should the notion of TAI bring?

To answer this question, let us focus on the role that this notion plays within the context of the European effort for regulating AI. In particular, let us look at the *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019). In this document,

[7] Note that here trustworthiness does not apply at the level of the system, but rather at the level of its outputs.

TAI is first of all characterized as grounded in four ethical principles (pp. 12–13). For starters, the principle of *respect for human autonomy* should be complied with, according to which "humans interacting with AI systems must be able to keep full and effective self-determination over themselves". The *prevention of harm* is another pillar: AI systems "should neither cause nor exacerbate harm or otherwise adversely affect human beings". The third principle, the principle of *fairness*, clearly states that "the development, deployment and use of AI systems must be fair". Finally, the *principle of explicability* should be respected: "processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions—to the extent possible—explainable to those directly and indirectly affected".[8]

When it comes to the practical realization of TAI, these principles are so to say translated into different practical "requirements": (i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency; (v) diversity, nondiscrimination and fairness; (vi) societal and environmental wellbeing; (vii) accountability. Note that we do not wish to extrapolate a clear-cut definition of TAI from these requirements, if only for the fact that the list is not meant to be exhaustive. However, we believe that the *Guidelines* provide precious insights into the advantages the notion of TAI is supposed to bring. In particular, they show that aspects related to algorithmic performance, such as accuracy and resilience to attacks, are not the full story when it comes to the design, assessment and evaluation of an AI system. In fact, they are subsumed unto the requirement of "technical robustness and safety", which is just one of the different requirements for TAI. In our view, this is exactly the point of having the notion of TAI on the table: it allows us to go beyond mere considerations of algorithmic performance.[9]

Let us be clear: aspects such as accuracy and robustness are pivotal for determining whether an artificial system is trustworthy. As a matter of fact, they are crucial when it comes to evaluating reliability, which we have seen is an indispensable component of trustworthiness. Accordingly, when deciding whether a given AI system should be trusted, our first considerations arguably concern its reliability. An example from real-world applications of AI technologies might help see this point. In recent years, AI algorithms have been increasingly used in medical contexts. Among other things, an interesting and promising application of AI techniques comes from oncological imaging. Let us consider systems based on deep learning (DL) for the detection of skin cancers. A crucial component of these systems' trustworthiness is their reliability in telling apart cancerous skin lesions from benign ones. Even though these technologies are relatively new and further research will arguably make them more efficient and accurate, results are already outstanding (Esteva et al. 2017; Soenksen et al. 2021). In fact, if one considers these systems' performances in terms of sensitivity and specificity based on their result in testing phases, they are perfectly comparable with and in many cases outperform their human counterparts.

This, however, is only half of the story. In addition to reliability, aspects such as avoidance of discrimination and opacity, representativity, and attribution of responsibility play a central role in dealing with AI systems. Our algorithm for skin cancer detection might have worked brilliantly in the testing phase. However, as widely acknowledged, algorithms trained on large datasets can easily incorporate bias and exacerbate discrimination. If training and testing data were under-representative of some population P, then using the algorithm for patients from P will be risky. Unfortunately, this is not a conjecture, and some algorithms used for the detection of skin cancers have proven to perform significantly worse on dark skin tones (Adamson and Smith 2018; Daneshjou et al. 2022). What is more, and less related to accuracy, these algorithms are riddled with problems of opacity. On the one hand, their DL architectures provide significant advantages in terms of performance. On the other hand, DL algorithms are notoriously opaque—as opposed to transparent—in the sense that their functioning is highly inscrutable (Topol 2019; Zerilli 2022). When the algorithm provides a certain prediction, it is virtually impossible to reconstruct the process that led to the system's output. We can clearly see why this is problematic by keeping considering medical contexts, even if the question generalizes. If the algorithm's diagnosis turns out to be wrong, who is responsible? A related point has to do with human oversight: to what extent should treatment plans be decided solely on the basis of the diagnostic algorithm's output?

The point is that AI systems are not ethically neutral, and value-laden choices are made at several stages in the making and use of algorithms (see Biddle 2022). We have seen that AI systems can exacerbate discrimination as a result of biased and nonrepresentative training and testing data. In addition, unfair choices can be made in the algorithm's design phase. In supervised learning, for instance, the programmer's biases could be directly inherited by the algorithm when the training labels are selected or the number and kinds of outputs are set (e.g., binary *versus* nonbinary gender taxonomies). What is more, as in the case of medical DL-based systems, trade-offs between algorithmic accuracy and explainability often need to be made.

[8] On the link between trust in AI and explainability, see, among others, Papagni et al. (2022).

[9] See also Russo et al. (2023) on the importance of combining epistemic and ethical aspects in the design, use, and assessment of AI systems.

When it comes to accommodating such value-laden aspects, the notion of reliability shows its limitations. The point is that we do not only want AI systems to be accurate, especially if accuracy is assessed during the algorithm's testing phase and not in real-world scenarios where the algorithm may encounter classes of inputs that were not sufficiently represented in training and testing data. We also want them to produce fair outcomes and to be reasonably transparent, and when transparency can hardly be achieved we want them to be constantly subject to assessment in their distribution and use.[10] In our view, these aspects of the design and use of AI algorithms should be captured by a different notion that goes beyond mere reliability and merely epistemic aspects. The notion of TAI seems to be suited for this scope, for it explicitly encompasses reliability and at the same time allows to capture ethically relevant elements.

If keeping together reliability and ethical aspects is the point of having the notion of TAI, as we contend, it is easy to see the limitations of the views on TAI presented above. On the one hand, if we build upon motivational accounts, then we are prevented from using the notion of TAI. In fact, according to these accounts, we should "abandon the 'trustworthy AI' paradigm as it is too fraught with problems, replacing it with the reliable AI approach" (Ryan 2020, p. 2765). On the other hand, opting for a purely epistemic account of TAI creates more problems than it solves. Aspects such as fairness and respect for human autonomy, which we have deemed essential and motivating the very use of the notion of TAI, are not only epistemic in nature, and therefore would not be captured by a purely epistemic understanding of trustworthiness. In both cases, although for different reasons, it all comes down to reliability. And unfortunately, reliability is not enough.

Taking stock, in this section we have presented and discussed the main positions in the debate on the notion of TAI. Both of them, we showed, fail to allow for a meaningful use of this notion and make its introduction largely useless. We have argued that a different strategy should be pursued and that the notion of TAI should be primarily discussed having in mind its use and the reasons behind its introduction. In particular, we have argued that the notion of TAI should be maintained to capture the nonepistemic and value-laden aspects of the design and use of AI systems.

5 A tale of two notions: trust in humans and AI systems

5.1 Trusting humans, trusting AI

So far, we have insisted on the importance of the notion of TAI. In this section, we wish to spell out in more detail some aspects of our proposal. In particular, we aim at making explicit the relationship between the notions of trust in AI systems and trust in humans – for brevity, we will refer to them respectively as H – AI (human – AI) trust and H – H (human – human) trust. We argue that H – AI and H – H trust are two distinct notions, held together by a common conceptual core.

Let us start with the differences. In the previous section, we have argued that, in addition to reliability, H – AI trust depends on features of AI systems that are relevant from an ethical and value-oriented perspective. Most notably, these features include respect for human autonomy and algorithmic transparency and fairness. Although allowing for a smooth application of a not purely epistemic notion of trust (and trustworthiness) to the field of AI and its products, this calls for some clarifications. As a matter of fact, regardless of how H – H trust is exactly spelt out, it seems fair to assume that its "extra factor" complementing reliability has little to do – if not indirectly – with human autonomy in human – machine interactions, as well as with algorithmic transparency and fairness. Again, the details vary depending on the specific accounts that one considers. However, as we have seen, views of H – H trust that do not exclusively consider the epistemic dimension typically focus on moral obligations, motivations and interests.

This being the situation, one might wonder what place the notion of H – AI trust occupies vis-à-vis H – H trust. As far as we can see, there are at least two choices if we wish to maintain the notion of trust in its application to AI and we are not content with a purely epistemic account of H – AI trust. A first possibility would be to maintain a single notion of trust but making it so general that it applies smoothly to both H – H and H – AI interactions. Otherwise, we can allow for the existence of a notion of trust (and trustworthiness) that is specific to AI [11]. The first option arguably scores

[10] The list is not meant to be exhaustive. Another interesting aspect, for example, has to do with the way AI systems are communicated to avoid AI anthropomorphism (see n. 4). Interestingly, the *Ethics Guidelines* explicitly state that "AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system", and that "the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand" (AI HLEG 2019, p. 18). In our view, the compliance with these requirements – that, let us note it, do not exclusively concern the design phase – plays an important role in increasing the levels of (well-founded) human trust in these systems.

[11] Note that both these options presuppose that the notions of trust and trustworthiness are used *literally* in their application to AI systems. However, a different strategy could be tried, involving understanding talk of trust in AI systems in metaphorical terms. Interestingly, this would be consistent with scientific practice in the field of AI, that makes extensive use of metaphors (Murray-Rust et al. 2022). We maintain an approach that allows us to literally apply trust and trustworthiness to AI systems, but we are aware that nonliterality

better in terms of conceptual simplicity, since there is just one notion instead of two. This simplicity, however, comes at the cost of threatening the very usefulness of the notion of trust, for the relevant and distinctive features of both H – H and H – AI interactions may be lost in the generalization. Based on these considerations, we deem it preferable to sacrifice conceptual simplicity in favour of two notions that preserve the specificities of different trust-based relations.

Needless to say, there must be some commonalities between them. We are not conceiving trust-based H – AI interactions as *radically* different from H – H ones. On the contrary, we hold that there is a *conceptual core* of trust that is shared by H – H and H – AI trust and is constituted by elements that are common to both. Identifying this conceptual core is pivotal for our proposal. Otherwise, among other things, it would make little sense to insist on the importance of the notion of *trustworthy* AI. If we did not want to put in prominence the fact that H – H and H – AI trust share important and distinctive elements, then we could just maintain the talk of trust for H – H interactions and use a completely different notion for AI systems. Although this is an open possibility, we believe that there are indeed some common elements that motivate the application of the notion of trust to AI systems. In what follows, we focus on the three elements that strike us as more important. The first two have largely been discussed in the previous sections, and we will just touch upon them. The third, instead, deserves a little more discussion.

The first common element we wish to highlight is that, just like trustworthy humans, trustworthy AI systems are first of all reliable. We have seen that accounts of H – H trust typically take reliance to be an indispensable component of trust. In the same way, we take for granted that TAI has to be first of all reliable AI – that is, robust AI that produces accurate results.

The second (related) element that is common to H – H and H – AI trust is that reliability is not enough. Under some prominent accounts of H – H trust, a nonepistemic extra element has to be present. With respect to AI systems, this point has been extensively addressed in Sect. 4, where we have argued that, just like in H – H trust-based relations, trust in H – AI interactions depends upon value-laden and ethically relevant aspects. Again, there are important differences between H – H and H – AI trust when it comes to the way these aspects are realized. Besides the differences, however, the point remains that reliably performing the task delegated by the trustor is not enough. In both cases, trustworthiness is supposed to work as a form of nonepistemic guarantee for the trustor.

But why is this guarantee necessary? This question brings us directly to the third element we wish to highlight. In both philosophy and sociology, H – H trust has often been conceived as occurring in contexts of uncertainty[12] and risk in which the trustor is vulnerable (Baier 1986; Hardin 2002; Luhmann 1979). If I trust my colleague Jessie to hold an old and rickety ladder while I climb it, I am clearly in a situation of vulnerability. This is not (only) due to the ladder's instability. I am also exposed to the possibility that Jessie gets distracted and lets go, or worse to the possibility that Jessie purposely lets me fall. As a matter of fact, I cannot be sure in advance that Jessie will do what asked to do. Needless to say, before climbing the ladder I will do my best to make sure that the person I asked to help is trustworthy – that is, that Jessie is not inattentive or willing to hurt me. Still, there is an element of risk given by the fact that I cannot be absolutely sure in advance whether Jessie will be a good trustee and will correctly perform the delegated task.

In all of this, the nonepistemic component of trust provides me with some form of guarantee: among other things, Jessie has the *moral obligation* not to let me fall on purpose. True, this does not change the fact that I am vulnerable: if the ladder breaks under my weight I get hurt. Nor does it dissolve the component of risk, for Jesse could still get distracted or betray me, choosing to disregard moral obligations. However, the relation of trust provides me with an acceptable way to deal with a risky situation in which I am vulnerable.

Mutatis mutandis, these considerations apply to AI systems as well. Let us consider the element of risk. Nowadays, AI systems are used on a regular basis in different critical contexts. We have already discussed the case of medical AI systems, and no explanation is needed for the claim that these systems operate in contexts of vulnerability. In addition, we can also think about the use of AI systems in courts. Here, one of the standard references is COMPAS, an AI system used in the U.S. to evaluate the likelihood of defendants' recidivism. With COMPAS, both false positives and false negatives have critical consequences. An overestimation of the risk of recidivism can result in excessively harsh

Footnote 11 (continued)

is an alternative option. In fact, if we understand talk of TAI as an instance of metaphorical language, the conceptual–nonsense objection loses its strength. To make talk of TAI legitimate, we would not need to grant that AI systems have motivations and moral obligations, nor should we excessively weaken the notion of trust. We could simply maintain that AI systems share some relevant features of trustworthy entities, grounding the felicitous use and understanding of the metaphor, although it is literally false that they can be trustworthy. Interestingly, these features would arguably be the same we are now going to list to identify what we refer to as the conceptual core of H – H and H – AI trust.

[12] Here, uncertainty is intuitively understood as generally opposed to certainty about the outcomes.

sentences, whereas an underestimation could in principle be harmful to society.

The list of critical contexts could go on for long, including war scenarios and financial markets. The point is that AI systems are increasingly employed in situations in which errors and misuse can have significant implications for human safety and well-being.[13] To make things worse, especially when machine learning techniques are employed, AI systems can be extremely opaque and we often ignore the process that led a given algorithm to provide a certain output. What is more, it is often impossible to check the data on which the algorithm has been trained and which heavily determine the algorithm's behaviour – just to give an idea, OpenAI's language model GPT3 was trained on 570 GB of filtered plaintext (Brown et al. 2020).

Faced with these problems, we resort to algorithms' trustworthiness, understood in the terms discussed in Sect. 4. In light of AI systems' intrinsic complexity and potential risks, we seek fair and possibly transparent AI algorithms, whose use is compatible with the respect of our autonomy and accompanied by human oversight. Again, this does not dissolve the problem: we are still vulnerable in a context of risk. However, the fact that the AI systems we use embed values and ethical constraints should provide us with some guarantee that does not reduce to the fact that the algorithms in question have performed well in their testing phase and previous use.

To sum up, we have identified three elements of trust and trustworthiness that are common to H – H and H – AI interactions: (i) reliability is the basis for trust; however, (ii) reliability is not enough, for the notion of trust is also grounded in an ethical dimension; finally, (iii) trust and trustworthiness provides us with a nonepistemic guarantee in contexts of vulnerability and risk. Identifying these elements allows us to maintain that H – H and H – AI trust are two distinct notions that nonetheless share a conceptual core and motivates our use of *trust* – and not some other notion – in applications to AI systems.

5.2 Possible objections

The distinction between H – H and H – AI trust prevents two possible objections from applying to our proposal. Let us start with the first one, which we may call the *conceptual stretch* objection. In a nutshell, one could argue that our notion of TAI builds upon a suspiciously *sui generis* understanding of trust, for our view on the ethical component of H – AI trust is grounded in aspects that are hardly present in H – H trust (algorithmic fairness, transparency, and so on). Building upon this point, one could contend that all we are doing is unduly stretching the notion of trust. While this would allow us to include AI systems in the range of trustable and possibly trustworthy entities, the resulting notion of trust would be too loose.

The second objection is the familiar conceptual nonsense objection. In Sect. 4, we have insisted on the importance of the ethical and value-laden aspects of TAI. Based on this, one might object that it is not clear how exactly an understanding of TAI such as the one we propose would not be subject to the conceptual nonsense objection. After all, we are willing to apply a not fully epistemic notion of trust to artefacts that lack the paradigmatically human features that are so crucial to H – H trust.

As far as we see, these two possible objections are motivated by a common assumption, namely that the notion of H – AI trust should be uncompromisingly modelled on H – H trust. Referring to noninterpersonal forms of trust, such as trust in governments, science, robots, and so on, McLeod (2021) notes that "most would agree that these forms of 'trust' are coherent only if they share important features of (i.e., can be modelled on) interpersonal trust". Narrowing it down to the literature on TAI, for instance, Ferrario et al. (2021) argue that "we shall strive, as much as possible, to identify a meaningful concept of trust that is applicable to human – human and human – AI relations". Under this methodological assumption, the notion of TAI as we conceive of it – and as it is used in the European *Ethics Guidelines* – is clearly problematic, for it explicitly revolves around aspects and requirements that can hardly be relevant for H – H interactions.

Here, however, we have explicitly rejected such an assumption, allowing for two distinct notions of trust that result in a difference in the aspects that contribute to trustworthiness. The first objection is thereby overcome: in our proposal, the *sui generis* character of trust in its application to AI systems does not come as an unwanted consequence. On the contrary, it is due to the fact that H – AI trust is indeed distinct from the kind of trust we are more familiar with, that is H – H trust. In other words, there is no conceptual stretch, just a legitimate and fruitful conceptual differentiation.

This differentiation also makes the conceptual nonsense objection innocuous. As a matter of fact, the crucial point of the objection is that paradigmatically human traits such as motivations and respect for moral obligations would be required for trustworthiness, and AI systems lack these traits. This objection, however, builds upon the identification of trust with H – H trust. In so far as we leave room for a notion of trust that is specific to AI systems, nothing prevents us from modelling it in a way that puts in prominence value-laden aspects of AI systems without requiring

[13] Nowotny (2021) provides a picture of the current status of AI and puts in prominence the uncertain character of the context in which these systems are integrated.

that AI systems themselves display paradigmatically human features. Again, this move is not a terminological trick. We are not using the label "trust" for referring to something that has nothing to do with H – H trust. On the contrary, we have identified a minimal core of trust that is common to both H – H and H – AI forms of trust, even if these depend on different factors (i.e., goodwill and moral obligations *versus* features such as transparency and fairness). However, the fact that they depend on different factors – and more precisely the fact that AI systems do not have to behave out of goodwill or possess moral obligations to be trustworthy – makes the conceptual nonsense objection ineffective.

Note that allowing for a distinction between H – H and H – AI trust should not pave the way for an uncontrolled bloating of the notions of trust and trustworthiness. That is, the fact that we can conceive these notions in a way that is specifically tailored to AI systems should not encourage to do the same with all technological artifact, thereby having different concepts for human – elevator trust, human – thermostat trust, and so on. As we have seen, AI systems are nowadays employed in critical and high-risk contexts, from healthcare to courts and working environments, where human decision making is increasingly substituted by AI-powered tools. Although other technological artifacts are also used in these contexts, AI systems are increasingly delegated with – or in any case decisive to – tasks and decisions having a huge impact on human lives. On top of that, AI systems stand apart from other artifacts in that their working – just like human behaviour – is distinctively characterized by high degrees of autonomy (Fossa et al. 2022), in a way that other artifacts are not, and often operate largely independently of human guidance. What is more, they do it in a way that is often opaque, without receiving explicit instructions on how inputs should be connected to outputs. Accordingly, just like in H – H trust, H – AI trust occurs in contexts in which the trustee's behaviour is uncertain, and trust provides the trustor with an extra guarantee that, whatever will be the trustee's behaviour, it will be bounded by ethical requirements. It is doubtful that trust would have the same role to play in the case of other artifacts, whose working could be largely dependent on human intervention and whose behaviour may be determined by specific instructions or components.

The final objection we wish to anticipate addresses the question of responsibility. The objection goes as follows: if the notions of trust and trustworthiness are modified so as to be applicable to AI systems, we risk removing responsibility from humans. Why not stretch instead the notion of reliance to include moral elements? This way, we could keep using the notion of reliance for AI systems and we would avoid the potential problems involved in the discourse on TAI. In other words, reliance should be revised to include an ethical alignment, and there would be no issue stemming from the applications of "trust" and "trustworthiness" to artificial agents. Among other things, this move may require the incorporation of ethical elements into the design and evaluation of technical artefacts, such as in the case of Value Sensitive Design (van den Hoven 2013). Now, although important for the anticipation of moral considerations already at the design level, Value Sensitive Design presents some possible limitations. It is not the place here to discuss all of them, but one of them plays an important role when considering the above objection. The point is that, by considering moral values in terms of design constraints, we run the risk of trying to solve ethical issues exclusively by technical means. Unfortunately, these considerations backfire on the objection we are considering. As a matter of fact, including ethical alignment in the notion of reliance raises a concrete risk of removing responsibility from humans and transforming it into something that can be mostly managed at the design level.

More generally, stretching the notions of reliance and reliability so as to include ethical desiderata, thereby avoiding talk of trust in AI systems, may not be convenient from a conceptual point of view. To begin, a strategy should be found to clearly distinguish reliance and reliability as typically understood in the literature (see §2) and their stretched-out counterparts encompassing the ethical component. Even granting that this distinction can be smoothly drawn, something would be still missing. Our attempt to identify a notion of trust that is specific and applicable to AI systems is motivated by the fact that the use of these systems typically takes place in contexts of risk in which the trustor is vulnerable. As we have seen, risk and vulnerability have systematically been associated with the notion of trust – in fact, it is not clear whether trust could occur in no-risk contexts in which the potential trustor is not vulnerable. These aspects, however, can hardly be captured by the notion of reliance.

6 Conclusion

The notion of TAI has become increasingly important in the debate on AI. In this paper, we have considered the main views in the philosophical literature on TAI, and we have argued that they fail to allow for a meaningful and productive use of this notion. We have insisted on the importance of a notion of TAI that captures both the epistemic and the non-epistemic dimensions of the design and use of AI systems. Moreover, by explicitly differentiating the notions of H – H and H – AI trust, we have provided a conceptual framework for talking about TAI without the risk of overly stretching the concept of trust or making categorical mistakes.

We have insisted on the ethical aspects of TAI and we have referred to some features that trustworthy AI systems should possess, but we have provided no exhaustive list of determinants of H – AI trust. In fact, one could even call into question the determinant-based approach, prevailing in the literature. Addressing these questions is crucial for both the philosophical debate on TAI and the attempt to provide an ethics-based regulation for AI. By clarifying the scope of the notion of TAI and providing the conceptual framework for its use, we hope we have taken a step in the right direction.

Acknowledgements We would like to thank Fabio Fossa, who provided us with extensive feedback on the manuscript. This work was supported by the Italian Ministry of University and Research under the PRIN Scheme (Project no. 2020SSKZ7R). Mattia Petrolo acknowledges the financial support of the FCT – Fundação para a Ciência e a Tecnologia (2022.08338.CEECIND; R&D Unit Grants UIDB/00678/2020 and UIDP/00678/2020). Viola Schiaffonati acknowledges the PNRR-PE-AI FAIR-NextGeneration EU program. Daniele Chiffi acknowledges the PNRR-RETURN-NextGeneration EU program.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adamson AS, Smith A (2018) Machine learning and health care disparities in dermatology. JAMA Dermatol 154(11):1247–1248. https://doi.org/10.1001/jamadermatol.2018.2348

Anderson M, Anderson S (eds) (2011) Machine Ethics. Cambridge University Press, Cambridge

Araujo T, Helberger N, Kruikemeier S et al (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc 35:611–623. https://doi.org/10.1007/s00146-019-00931-w

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

Biddle JB (2022) On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. Can J Philos 52(3):321–341. https://doi.org/10.1017/can.2020.27

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Amodei D. (2020). Language Models are Few-Shot Learners. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 1877–1901

Buechner J, Simon J, Tavani HT. (2014). Re-Thinking Trust and Trustworthiness in Digital Environments. In Buchanan E. et al. (Eds.), Autonomous Technologies: Philosophical Issues, Practical Solutions, Human Nature. Proceedings of the Tenth International Conference on Computer Ethics Philosophical Enquiry, INSEIT, 65–79

Coeckelbergh M (2012) Can We Trust Robots? Ethics Inf Technol 14(1):53–60. https://doi.org/10.1007/s10676-011-9279-1

Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, Ko J, Swetter SM, Bailey EE, Gevaert O, Mukherjee P, Phung M, Yekrang K, Fong B, Sahasrabudhe R, Allerup JAC, Okata-Karigane U, Zou J, Chiou A (2022) Disparities in dermatology AI performance on a diverse, curated clinical image set. Sci Adv 8(32):6147. https://doi.org/10.1126/sciadv.abq6147

DeCamp M, Tilburt JC (2019) Why we cannot trust artificial intelligence in medicine. Lancet Digital Health 1(8):e390. https://doi.org/10.1016/S2589-7500(19)30197-9

Durán JM, Formanek N (2018) Grounds for trust: essential epistemic opacity and computational reliabilism. Mind Mach 28(4):645–666. https://doi.org/10.1007/s11023-018-9481-6

Durán JM, Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. J Med Ethics 47:329–335. https://doi.org/10.1136/medethics-2020-106820

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:7639. https://doi.org/10.1038/nature21056

European Commission (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence—Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain Union legislative acts. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

Ferrario A, Loi M, Viganò E (2020) In AI we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. Phil Technol 33(3):523–539. https://doi.org/10.1007/s13347-019-00378-3

Ferrario A, Loi M, Viganò E (2021) Trust does not need to be human: It is possible to trust medical AI. J Med Ethics 47(6):437–438. https://doi.org/10.1136/medethics-2020-106922

Fossa F (2019) «I don't trust you, you faker!» on trust, reliance, and artificial agency. TESOL J 39(1):63–80. https://doi.org/10.4454/teoria.v39i1.57

Fossa F, Chiffi D, De Florio C (2022) A Conceptual Characterization of Autonomy in the Philosophy of Robotics. In: Riva G, Marchetti A (eds) Humane Robotics. A Multidisciplinary Approach Towards the Development of Humane-Centred Technologies. Vita e Pensiero, Milano

Gambetta D (1988) Can We Trust Trust? In: Gambetta D (ed) Trust: Making and Breaking Cooperative Relations. Blackwell, Oxford, pp 213–237

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

Grodzinsky F, Miller K, Wolf MJ (2020) Trust in artificial agents. In: Simon J (ed) The Routledge Handbook of Trust and Philosophy. Routledge, New York, pp 298–312

Hardin R (2002) Trust and Trustworthiness. Russell Sage Foundation, New York

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

AI HLEG (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hoff KA, Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434

Hunyadi M (2019) Artificial Moral Agents. Really? In: Laumond J-P, Danblon E, Pieters C (eds) Wording robotics: discourses and representations on robotics. Springer International Publishing, Cham

Klein HA, Lin M-H, Miller NL, Militello LG, Lyons JB, Finkeldey JG (2019) Trust across culture and context. J Cognitive Eng Decision Making 13(1):10–29. https://doi.org/10.1177/1555343418810936

Luhmann N (1979) Trust and Power: Two Works. Wiley, Chichester

Lünich M, Kieslich K (2022) Exploring the roles of trust and social group preference on the legitimacy of algorithmic decision-making vs. human decision-making for allocating COVID-19 vaccinations. AI Soc. https://doi.org/10.1007/s00146-022-01412-3

McLeod C. (2021). Trust. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entriesrust/

Metzinger T. (2019). EU guidelines: Ethics washing made in Europe. Der Tagesspiegel Online. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Mökander J, Juneja P, Watson DS, Floridi L (2022) The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What can they learn from each other? Mind Mach 32(4):751–758. https://doi.org/10.1007/s11023-022-09612-y

Mollering G (2006) Trust: Reason, Routine, Reflexivity. Bingley, Emerald Group

Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21

Murray-Rust, D. S., Nicenboim, I., & Lockton, D. (2022). Metaphors for Designers Working with AI. In DRS Conference Proceedings 2022 (RS Biennial Conference Series). https://doi.org/10.21606/drs.2022.667

Nguyen CT (2022) Trust as an Unquestioning Attitude. In: Gendler TS, Hawthorne J, Chung J (eds) Oxford Studies in Epistemology, 7. Oxford University Press, Oxford

Nickel PJ (2007) Trust and obligation-ascription. Ethical Theory Moral Pract 10(3):309–319. https://doi.org/10.1007/s10677-007-9069-3

Nickel PJ, Franssen M, Kroes P (2010) Can we make sense of the notion of trustworthy technology? Knowl Technol Policy 23(3):429–444. https://doi.org/10.1007/s12130-010-9124-6

Nowotny H (2021) In AI we trust: power, illusion and control of predictive algorithms. Polity, Cambridge

Papagni G, de Pagter J, Zafari S et al (2022) Artificial agents' explainability to support trust: considerations on timing and context. AI & Soc. https://doi.org/10.1007/s00146-022-01462-7

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

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

Simon J (ed) (2020) The Routledge Handbook of Trust and Philosophy. Routledge, New York

Skjuve M, Følstad A, Fostervold KI, Brandtzaeg PB (2021) My chatbot companion-a study of human-chatbot relationships. Int J Hum Comput Stud 149:102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Soenksen LR, Kassis T, Conover ST, Marti-Fuster B, Birkenfeld JS, Tucker-Schwartz J, Naseem A, Stavert RR, Kim CC, Senna MM, Avilés-Izquierdo J, Collins JJ, Barzilay R, Gray ML (2021) Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Sci Transl Med 13(581):3652. https://doi.org/10.1126/scitranslmed.abb3652

Starke G, Brule R, Elger BS, Haselager P (2022) Intentional machines: a defence of trust in medical artificial intelligence. Bioethics 36(2):154–161. https://doi.org/10.1111/bioe.12891

Taddeo M (2009) Defining trust and E-trust: from old theories to new problems. Int J Technol Human Interact 5(2):23–35. https://doi.org/10.4018/jthi.2009040102

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:1. https://doi.org/10.1038/s41591-018-0300-7

van den Hoven J (2013) Value sensitive design and responsible innovation. In: Owen R, Bessant J, Heintz H (eds) Responsible innovation: managing the responsible emergence of science and innovation in society. Wiley, London, pp 75–83

Zerilli J (2022) Explaining machine learning decisions. Philos Sci 89(1):1–19. https://doi.org/10.1017/psa.2021.13

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