Trust and Trustworthiness in AI

Durán, Juan Manuel; Pozzi, Giorgia

DOI
10.1007/s13347-025-00843-2

Publication date
2025

Document Version
Final published version

Published in
Philosophy and Technology

Philosophy & Technology (2025) 38:16
https://doi.org/10.1007/s13347-025-00843-2

RESEARCH ARTICLE

Trust and Trustworthiness in AI

Juan Manuel Durán[1] · Giorgia Pozzi[1]

Received: 23 July 2024 / Accepted: 17 January 2025

Abstract
Achieving trustworthy AI is increasingly considered an essential desideratum to integrate AI systems into sensitive societal fields, such as criminal justice, finance, medicine, and healthcare, among others. For this reason, it is important to spell out clearly its characteristics, merits, and shortcomings. This article is the first survey in the specialized literature that maps out the philosophical landscape surrounding trust and trustworthiness in AI. To achieve our goals, we proceed as follows. We start by discussing philosophical positions on trust and trustworthiness, focusing on interpersonal accounts of trust. This allows us to explain why trust, in its most general terms, is to be understood as reliance plus some "extra factor". We then turn to the first part of the definition provided, i.e., reliance, and analyze two opposing approaches to establishing AI systems' reliability. On the one hand, we consider *transparency* and, on the other, *computational reliabilism*. Subsequently, we focus on debates revolving around the "extra factor". To this end, we consider viewpoints that most actively resist the possibility and desirability of trusting AI systems before turning to the analysis of the most prominent advocates of it. Finally, we take up the main conclusions of the previous sections and briefly point at issues that remain open and need further attention.

Keywords Trustworthy AI · Trust · Extra factor · Transparency · Computational reliabilism

1 Introduction

Establishing AI systems' trustworthiness is increasingly considered fundamental for their integration into society. This holds particularly true in human-sensitive domains such as medicine, healthcare, employment, government, energy, criminal justice, and security. The general principles for trustworthy AI outlined by the EU

✉ Juan Manuel Durán
j.m.duran@tudelft.nl

[1] Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft 2628 BX, The Netherlands

Published online: 04 February 2025

Commission (2019), echoed throughout the specialized literature (e.g., Li et al., 2023; Kaur et al., 2022), advocate for caution and the pursuit of robust solutions. Many technical solutions are available today that aim to fortify our trust in AI and ensure their trustworthiness (e.g., Cho, 2019). But what makes an AI trustworthy? Why should we trust its output and behavior? Does it come down to merely scrutinizing the algorithm's patterns, or is there more to it?

To illustrate the interplay between trust and trustworthiness and set the stage for the goals of this paper, consider a case of interpersonal trust that is easily relatable. We place our trust in physicians because they have undergone medical school, acquired the knowledge of medicine, and possess the ability to apply medical care in specific situations. Philosophically, this is referred to as the *reliance* on the trustee, i.e., the physician. Reliance, in this context, is an epistemic term, signifying a property that something or someone upholds for being *trustworthy*. We rely on the physician's competence based on having the right education. We rely on the bus because it is always on time. But is reliance alone sufficient for trust? Can we simply say we trust the physician because they went to medical school?

Our *trust* in the physician extends beyond the expectation that they will prescribe the right medicines and make accurate diagnoses. By trusting, we also hold a normative expectation that the physician will do the right thing. For instance, we expect them to act in our best interest, in accordance with the biomedical principles of beneficence and non-maleficience. In other words, trust places a moral demand on the physician to act in ways that surpass the mere value of their medical knowledge. It follows that reliance must be complemented with additional considerations to constitute genuine trust. This is where we introduce the ambivalent concept of an "extra factor," often taking the form of responsibility, commitment, and goodwill.

While the example above may apply to interpersonal trust between two humans (a trustor and a trustee, a physician and a patient), its relevance to AI systems is less clear. Consider the "extra factor," for instance. Can we demand responsibility from an AI, and if so, what would that entail? More provocatively, can we expect an AI to have our best interests "at heart"? These questions form the foundation of considerations about *trustworthy AI*, as defined by the EU Guidelines. This article aims to analyze the complexity of this issue by first distinguishing *trust* from *trustworthiness*, and then discussing the former as a two-part concept: reliance and the "extra factor."

With these goals in mind, we divide this article into three main sections. Section 2 briefly presents the philosophical literature on trust and trustworthiness. A key takeaway from this section is that studies on trust in AI can be analytically divided into two categories: reliance, which examines the conditions for scientifically valid outputs, and an "extra factor," which explores the moral motivations underlying trust (Hawley, 2019).

Section 3 focuses on the epistemic basis for relations of trust that later serve for securing trustworthy AI, that is, reliance. We treat reliance as the property of an algorithm of forming beliefs about the scientific validity of the outputs. Two main approaches emerge prominently: transparency, widely popular both in philosophical and data science circles, and computational reliabilism (CR), much less known but a major contender to transparency. In this section, we spell out the merits and shortcomings of both transparency and CR as two viable approaches to ensure the reliability of AI systems.

In Section 4, we turn to the second part of the definition of trust, focusing on discussions revolving around the "extra factor". Here, we address the fragmented philosophical debates on trust in AI in an attempt to bring some order to the discourse. To this end, we subdivide the debates between those who argue that trusting AI is *not possible* (or even undesirable) and those who maintain that trust in AI is *possible and much needed* (see Fig. 1). We approach these debates critically, highlighting their merits and shortcomings. This should contribute to the analysis of different positions regarding the "extra factor" concerning both the conceptual and normative possibility of genuinely trusting (and not merely relying on) AI systems.

Finally, in Section 5, we provide a brief summary of the main findings of this article, and we sketch some suggestions for further research revolving around trust, trustworthiness, and AI systems. A summary of the key concepts used in this entry, along with definitions and proposed relevant literature, can be found in Table 1.

2 The Multiple Dimensions of Trust and Trustworthiness

The concepts of trust and trustworthiness are ubiquitous in our daily lives, forming the bedrock of interpersonal relationships and societal dynamics. Despite this fact, it is remarkably difficult to define them in satisfactory terms that encapsulate their complexity and elucidate their fundamental features. Typically, we use these concepts across various contexts to govern interpersonal relationships and articulate our expectations from others. We express our present and future trust in our physicians and friends because they have shown to be trustworthy. We might, however, be less inclined to extend the same level of trust to a politician who has exhibited signs of untrustworthiness in the past. Similarly, establishing clear conceptual distinctions between what it means to trust that someone will fulfill a promise and merely hoping they will is far from straightforward.

These intricate issues deepen when AI systems become integral to decision-making contexts where the judicious allocation of trust holds significant importance. AI systems serve as mediators in trust relationships among different stakeholders, such as between doctors and patients, banks and loaners.[1] They are also positioned to be the direct recipients of our trust, as seen in recent cases in the judicial system

[1] Let us clarify what we mean when stating that AI systems often mediate trust relationships between humans. Consider, for example, NarxCare algorithms that are widely used in the USA to assess patients' eligibility for opioid medication by producing a risk score that predicts patients' probability of addiction or misuse of pain medication (Szalavitz, 2021). In these situations, the AI system *mediates* the trust that a physician puts (or withhold) in a patients' testimony because it provides actionable information about the patients' health status (Pozzi, 2023). Similar dynamics can occus if an AI system classifies an applicant as not eligible for a loan and, based on this prediction, this patient is distrusted in their aibility to pay a loan.

Table 1 Key concepts, definitions and proposed relevant literature

Key concepts	Definition(s)	Relevant Literature
Trust	Attitude of a trustor toward a trustee (feelings of betrayal upon its breach)	(McLeod, 2021)
Trustworthiness	Set of properties of the trustee (that give the trustor reasons to trust)	(Hardin, 2002)
Reliance	Expectation that the trustee will continue to perform as expected (disappointment but not betrayal upon failed reliance); in AI, justification that an AI produces scientifically valid outputs; epistemic attitude of humans toward AI systems	(Ryan, 2020), (Durán & Formanek, 2018), (Durán, forthcoming)
Transparency	Revealing hidden workings, causal connections, low-level mechanistic relations, and interdependencies within algorithm to enhance understanding and form beliefs about its output.	(Creel, 2020), (Datta et al., 2016), (Guidotti et al., 2019), (Watson & Floridi, 2021), (Von Eschenbach, 2021)
Computational Reliabilism	Accepting AI's epistemic opacity, dispositional account for belief formation through reliability indicators (system's technical robustness, compatibility with computer-based scientific practice, social construction of scientific beliefs)	(Durán & Formanek, 2018), (Durán, forthcoming), (Durán & Jongsma, 2021), (Ferrario et al., 2021)
Extra factor	● Expectation that the trustee will be favorably moved of the trustee (Jones) Expectations + motivations/goodwill of the trustee (Hardin) ● The trustor relies on the trustee being willing to do what they are entrusted with (McLeod) ● Commitment of the trustee (Hawley)	(Jones, 1996), (Hardin, 2002), (McLeod, 2021), (Hawley, 2014, 2019)
Agency-based account	AI systems have minimal agency; intentionality (minimal sense); AI systems are bearers of normative commitments	(Chen, 2021), (Lewis & Marsh, 2022), (Starke et al., 2022)
Rational choice account	Non normative, non-affective account; trust in AI amounts to relying on it without updating our beliefs (relying without monitoring)	(Ferrario et al., 2020), (Ferrario et al., 2021)
Discretionary account	Trust manifests in the discretionary authority attributed to AI; normative account (AI object of normative expectations, designers are the bearers of it)	(Nickel, 2022)
Reductive account	Trust not in AI directly but in AI its human designers and developers	(Sutrop, 2019)

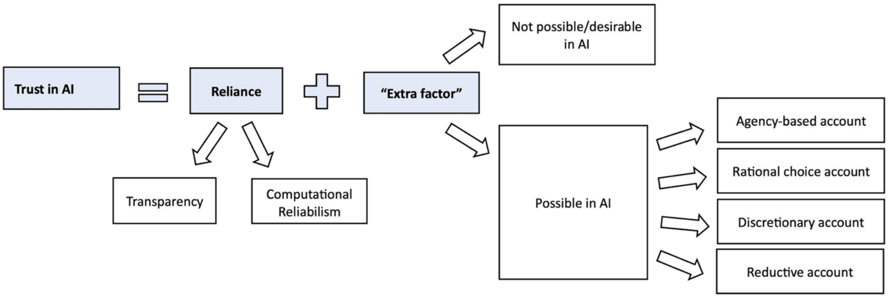

Fig. 1 Definition of trust in AI as reliance plus some "extra factor"

(e.g., COMPAS (Wexler, 2017). Similar to interpersonal trust, the users' trust in AI systems is foundational for their acceptance and successful integration into relevant social practices (Choung et al., 2022).

Philosophical inquiries into trust and trustworthiness focus on understanding the practical commitment between a trustor and a trustee. Conventionally, this involves the expectation that the trustee will fulfill the commitments made to the trustor or undertake actions deemed appropriate based on their expertise, training, or responsibility (Hawley, 2019). For instance, a trustworthy physician delivers accurate diagnoses, while a trustworthy friend safeguards shared secrets. Now, our trust in physicians also extends beyond their medical training, and the trust in our friends is not solely based on the fulfillment of their promises. Our trust in physicians is rooted in their commitment to our well-being (beneficence) and the prevention of harm (non-maleficence) or in assuming moral responsibility for their actions. Similarly, our trust in friends arises from their genuine affection and their willingness to refrain from deceiving us. This perspective underscores that trust involves the readiness of the trustor to put themselves in a situation of vulnerability, uncertainty, and risk (Lewis & Marsh, 2022).[2] It's worth noting that trust becomes relevant precisely in situations lacking full control, where delegation of a specific task to a trustee is necessary (McLeod, 2021).

Let us also note that we start with the premise that both the trustor and the trustee are individual human agents (e.g., ourselves, our physicians, our friends). Extending these results to a collective of agents (e.g., the International Panel for Climate Change) or, more abstractly, to institutions (e.g., the WHO or the government of The Netherlands) should not prove overly challenging. It is, in principle, appropriate to assert that an institution like the WHO is trustworthy based on its prioritization of global public health in its decisions regarding COVID-19.

[2] Here, our analysis is limited to instances of trust where a trustor delegates a specific task to a trustee with the aim of achieving a particular goal (*x* trusts *y* to do *z*) (McLeod, 2021). Trust, by its nature, is predominantly contextual. For instance, one might trust their physician for medical prescriptions but not for car repairs. Concepts like "generalized trust" (Hardin, 2002), such as trustworthiness as a character disposition or virtue, and broader notions of trust that extend beyond specific tasks or actions, will not be addressed in this article.

Problems arise, however, when the trustee is an instrument or, in the case of AI, a computational algorithm. To illustrate this, contrast our discussion with the feelings that arise when the practical commitment of trust is breached.[3] Instances of betrayal, deception, disappointment, and disgruntlement surface when we discover our physicians misdiagnosing us or our friend revealing a sworn secret. In cases similar to these, a breach of trust triggers genuine moral reactions due to the fact that the trustee fails to meet the normative or affective expectations we have of them. We expect from our physicians that they should act in our best interest and of our friends that they care so much about us to not reveal private information. In such instances, it is deemed appropriate to feel betrayed, and there is a rightful demand for an explanation regarding the failure to fulfill a specific commitment, or an apology is warranted. But it seems unfitting to feel disgruntled when a light bulb is not working or consider that the car has betrayed us when the engine does not start. There is a figurative use of trust applicable to these inanimate objects; for example, we trust that the lamp will light up as long as its basic functioning is unaltered, and we trust the car to start in the cold morning. Sometimes, we even say that the car is trustworthy or that such and such a company builds trustworthy cars. However, it needs to be clear that our relationship with these objects is one of reliance for specific purposes, not necessarily of trust. While it is appropriate to rely on the well-functioning of the lamp and the car, expecting loyalty from the former or demanding an apology from the latter would be inappropriate (Hawley, 2019, p. 2). A similar perspective seems to apply to computer-based algorithms.

The central issue here is that trust and trustworthiness have predominantly been conceptualized in *anthropomorphic* terms, characterized by their appeal to distinctly human and morally-laden emotions (such as betrayal, deception, intentionality, accountability, etc.). Approached in this manner, assertions about trust and trustworthy AI may seem inappropriate, unwarranted, and potentially misleading. As we will delve into shortly (section 4.1), this stance is firmly held by many philosophers working in this domain. The alternative involves formulating an account of trust and trustworthiness that explicitly incorporates AI as a significant component of its study. While current approaches exhibit notable shortcomings, there are compelling arguments that could guide us in gaining a more nuanced understanding of how to navigate these intricate issues (section 4.2).

[3] Let us point out that trust, distrust, and misplaced trust are conceptually distinguished attitudes. Distrust does not amount to a mere lack of trust since distrust entails a moral criticism that a lack of trust does not (Hawley, 2017). For instance, one may *distrust* a friend who has revealed a secret in the past. Here, distrust is appropriate because the friend has demonstrably not respected the commitment she made (i.e., to keep my secret). Conversely, a *lack of trust* can occur in situations in which neither trust nor distrust would be appropriate. For example, I neither trust nor distrust my physician with the repair of my car simply because neither attitude pertains to the task domain of my physician. As for *misplaced trust*, Nickel considers situations in which physicians perform defensive medicine, e.g., by over-prescribing medication (say, antibiotics for the common cold), due to the fear of breaching the trust patients put in them (Nickel, 2009). However, expecting a physician to prescribe antibiotics for a simple cold is not a measure of the physician's trustworthiness, and thus, as Nickel argues, trust in this case is misplaced (see also Hawley (2015). In this article, we focus exclusively on relations of trust.

Let us now briefly but explicitly articulate the philosophical distinction between trust and trustworthiness. Trust is considered an *attitude* that reflects the trustor's inclination to place trust in the trustee. This is why we say we trust our physician, signifying an attitude of confidence, belief, or a similar sentiment towards the physician. On the other hand, trustworthiness is a *property* intrinsic to the trustee. It represents what makes a trustee "demonstrably worthy of trust" (Sutrop, 2019). In this context, we label the physician as trustworthy because they have demonstrated that they are deserving of our trust.

Thus understood, trust and trustworthiness are distinct but related concepts, allowing for the possibility of trusting an untrustworthy person or entity and, vice versa, withholding trust even when the trustee is, in fact, trustworthy. Let us briefly consider both cases in turn.

Situations in which we trust the untrustworthy typically occur when we do not have much information about the trustee. For instance, consider the case of Zholia Alemi, who was found guilty of fraud for practicing as a psychiatrist for over 20 years without having acquired any medical qualifications (Bugel, 2023). Even though someone who falsifies a medical degree cannot be considered trustworthy, it is very likely that her patients trusted that she was a reliable and competent professional. The reason for this was a lack of information regarding the fact that she had not been to medical school. Knowing this information would have altered her patients' attitude of trust.

Instances in which trust is withheld, even though the trustee is trustworthy, often arise when the trustor holds biases that deflate the perceived trustworthiness of the trustee. Fricker's work on epistemic injustice underscores precisely this: individuals can fail to trust the trustworthy due to biases related to their interlocutor's social identity (e.g., biases related to gender, race, ability, socio-economic status) (Fricker, 2007). This phenomenon typicallyoccurs in interpersonal relationships due to implicit biases a person might hold, but they can also be fueled by explicitly discriminatory and stigmatizing public attitudes and statements. Notable instances of this phenomenon emerged among Trump supporters after hearing his controversial remarks about immigrants during his presidential campaign announcement speech on June 16, 2015. In that speech, he stated, ”When Mexico sends its people, they're not sending their best… They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people." (Phillips, 2017) This dubious and profoundly discriminatory statement most likely lead to failing to trust trustworthy individuals simply because they are the object of unfounded prejudices.

Now, while conceptually distinct, trust and trustworthiness are deeply interconnected in the sense that the presence of one entails the other. In other words, it is impossible to conceive (mis)trusting someone –or something– that lacks the property of being (un)trustworthy. This interconnection between trust and trustworthiness is pivotal in the debate over *trustworthy AI* –now understood in the general sense of the EU Guidelines.

Asserting that an AI system is trustworthy necessitates establishing the reliance of the system. It seems rather obvious that an AI with predictive accuracy for cancerous moles closer to 95% is deemed more reliable than one in the vicinity of 35%.

High predictive accuracy, however, is not the sole criterion for entrenching reliance on a given AI system. One might argue that explainability is the key property to this end. Likewise, one might request that the AI system possesses specific scientific merits that make it more (or less) reliable. In Section 3, we explore various options where the property of being reliable is elaborated and defended. Our focus centers on ongoing discussions on transparency and computational reliabilism, drawing insights from both philosophical and technical literature.

Trust, on the other hand, is a more complex issue for a comprehensive understanding of *trustworthy AI*. Recall that trust is an attitude pertaining to the trustor to be inclined to trust the trustee. As such, it requires not only some degree of reliance on the trustee but also "an extra factor" (Hawley, 2014, p. 5). Take again the case of trusting a physician. It is not enough to deem them trustworthy just given the right credentials and certificates. Proper relations of trust only surface when the physician shows to be responsible for our well-being or has our best interests at heart. This complexity of trust can also be illustrated with AI systems. Consider a Convolutional Neural Network (CNN) accurately predicting criminals based on facial traits (e.g., curvature of the mouth, distance between the eyes, etc.). While this CNN can be deemed reliable due to highly accurate predictions –it has been reported an estimate of 95% accuracy (Wu & Zhang, 2016)– it cannot be trusted in its outputs. If a judge were to sentence a person to prison based on the curvature of their nose, it would not only violate their rights and due process, neglecting the principles of fair and unbiased judgment, but also undermine justice, equality, human rights, and could lead to severe consequences such as wrongful imprisonment and perpetuation of systemic biases. At the same time, we cannot genuinely say that this DNN is responsible for its output, or it has the best intentions "at heart." Solving the issue of trust in AI is at the root of any comprehensive understanding of *trustworthy AI*. Philosophers recognize the difficulties of it, particularly pinning down the "extra factor."

What, then, is this "extra factor" exactly? Opinions among philosophers are divided. Some interpret it as a positive view of the *motives* of the trusted person. For instance, one might trust a physician because they have the right motives to look after one's health. However, defining what constitutes the "right motive" requires further clarification. Is it because physicians are bound to the Hippocratic Oath, or is it due to legal accountability? On the other hand, some consider the "extra factor" as a *reasonable expectation* on the trusted. Jones, for example, defines it as "the expectation that the one trusted will be directly and favorably moved by the thought that we are counting on her" (Jones, 1996, p. 1). Yet others, such as Hardin, combine *expectations with motives*, stating that "the truster's expectations of the trusted's behavior depend on assessments of certain motivations of the trusted" (Hardin, 2002, xix). McLeod points out that the extra factor "generally concerns why the trustor (i.e., the one trusting) would rely on the trustee to be willing to do what they are trusted to do" (McLeod, 2021). This latter perspective puts the focus on the *willingness* of the trustee.[4]

[4] For a critical review, see also Goldberg (2020).

Drawing from this literature, philosophers make efforts to accommodate AI. Within this context, two primary lines of argumentation surface. The first contends that the anthropomorphization of trust inherently rules out any possibility of trusting AI. That is, trust requires some form of responsibility, intentions, or normative commitment, none of which can be ascribed to algorithms. Section 4.1 presents and discusses the main proponents of this view. The second line of argumentation posits that trust in AI is indeed feasible; we just need to accept some assumptions and conditions. Section 4.2 discusses this possibility. Let us finally mention that our treatment intentionally simplifies various issues and, for instance, we will not explore the role that the trustor's prior beliefs might or might not play in establishing a relationship of trust with an AI system.

3 Reliance

It seems rather uncontroversial to say that we must secure an AI system's reliance before crediting our trust in it. After all, one would not trust a physician if they did not attend medical school. However, as uncontroversial as it might seem, it is far from clear how the reliance of an AI system can be established. In what follows, two theoretical frameworks are explored for establishing the general reliance of AI systems: *transparency* and *computational reliabilism*. Of specific interest is the application of AI in the scientific field.

Before we begin, two conceptual clarifications need our attention. First, reliance is not taken to be a property that AI systems have or fail to have. Rather, it comes in degrees. For instance, an AI system is reliable because it forms accurate beliefs most of the time; or its output is reliable because we managed to get some degree of transparency that warrant our beliefs. Second, we understand reliance as an epistemology by which we can justifiably state that an AI system is reliable or renders scientifically valid outputs.[5]

3.1 Transparency

Transparency is undeniably one of the most highly regarded methods for justifying our beliefs that the AI output is scientifically valid.[6] The underlying sentiment is

[5] We maintain a neutral stance regarding the precise definition of "a scientifically valid output." This concept can encompass various interpretations, such as being acceptable in terms of empirical predictions, formally correct, theoretically sound, and more. The specific criteria for scientific validity may vary depending on the context and the goals of the AI system.

[6] Transparency is a polysemous concept. For instance, transparency applies to the readiness of a company to share relevant information with their stakeholders (European Commission, 2019, p. 3), or of a government to disclose their plans. Thus understood, transparency amounts to a commitment to openly share information, processes, and decision-making with the public or its stakeholders. It involves clear communication, accessibility of relevant data, and a commitment to accountability. Transparency fosters trust by allowing external scrutiny, enabling informed decision-making, and demonstrating adherence to ethical and responsible practices. Here we exclusively consider it in its *epistemic* sense of justifying our belief in the output of an AI system.

genuine: when we can clearly understand how a system operates, we have reaons or supporting evidence to believe that its outputs hold scientific merit. As articulated by Guidotti et al., "[t]he availability of transparent machine-learning technologies would lead to a gain of trust and awareness on the fact that it is always possible to know the reasons for a decision or an event" (Guidotti et al., 2019, 93:2) In this respect, it's crucial to explore what transparency entails and how we can attain it.

The initial approach to defining transparency is to consider it as the opposite of opaque or "black box" algorithms, as suggested by Lipton (2018) and Creel (2020, p. 569. Footnote 2). In simpler terms, a transparent AI system is one that is not opaque. Unfortunately, this interpretation is not very illuminating, raising questions about what constitutes an opaque system and what exactly is meant by the opposite of opacity. Furthermore, it fails to recognize that opacity can take on different forms, including epistemic, methodological, and semantic opacity. Epistemic opacity, for instance, refers to the inherent cognitive limitations of humans to comprehensively understand and account for the state of a computer process, encompassing variables, system relations, and system status (see (Humphreys, 2009, p. 618) and (Durán & Formanek, 2018). Methodological opacity, on the other hand, concerns the coding practices and strategies used in the development of AI systems that are not always readily accessible to developers. These coding practices may involve complex algorithms or proprietary techniques that are not easily discernible (Burrell, 2016). Finally, semantic opacity relates to the difficulty in establishing a direct and meaningful representation between the AI system and real-world phenomena. This challenge arises from the abstract nature of AI algorithms, which might not always align perfectly with the complexities of the real world they seek to model or interact with (Humphreys, 2009, p. 619).

To illustrate the challenges in defining transparency in these terms, let's consider the opposite of epistemic opacity, which is *epistemic transparency*. In this context, transparency means the cognitive ability to comprehensively survey and account for variables, system relations, and other elements within the algorithm with the purpose of having reasons to believe in the scientific validity of the output. To demonstrate this interpretation, we can examine any Deep Neural Network (DNN). It is impossible for any human agent or group of agents to halt a DNN at a specific time *t* and assert full knowledge of the DNN's general state at that moment (e.g., which values have been instantiated for various variables). Similarly, predicting the DNN's next step at time *t + 1* (including computing the next step and determining which variables will be instantiated) or retroactively accounting for the past state of the DNN at *t-1* (e.g., identifying which variables were instantiated in the previous run) is exceptionally challenging. In summary, epistemic transparency implies having what could be presumed as complete cognitive access to the DNN at *t-1*, *t*, and *t + 1*, as well as the ability to provide meaningful insights about the algorithm. However, it is a well-established fact that achieving such comprehensive access is not cognitively possible for human agents, especially when complex AI systems like DNNs are involved.

The problem here is that opacity tends to be seen in absolute terms: algorithms are either opaque or not, with many of them exhibiting opacity on one or more levels (epistemic, methodological, semantic). In contrast, transparency is a concept that exists along a continuum involving degrees. It is, therefore, quite difficult to define one in terms of the opposition.

There is a more nuanced interpretation of transparency that has been articulated by Creel, who identifies three distinct forms or levels of it (Creel, 2020, p. 569):

1. *Functional Transparency*: This refers to having knowledge of how the algorithm as a whole functions and operates.
2. *Structural Transparency*: This involves knowledge of how the algorithm is implemented in code, essentially the coding details that make it work.
3. *Run Transparency*: This is concerned with knowledge of how the program actually operates in a specific instance, including the hardware and input data used during execution.

While it is useful to distinguish between these different sources of transparency, Creel's framework does not explicitly address how effectively each form of transparency can be achieved. This leaves room for the possibility that there may be multiple existing methods to attain each individual form of transparency, diverse degrees of transparency, and incompatibilities among methods (e.g., different approaches may prioritize one aspect of transparency over others or employ different techniques and trade-offs). This underscores the complexity and multifaceted nature of transparency in the context of AI and computational systems.

Perhaps the most widely accepted interpretation of transparency involves making visible the low-level mechanistic relations that underlie *how* an algorithm operates. This interpretation places significant emphasis on revealing the inner workings, causal connections, and interdependencies within the algorithm to enhance our understanding of its functioning and the outputs it produces. Following the literature, let us call it "opening the black box."

Now, there are several ways to advocate for opening the black box. One is to consider uncovering the hidden causal structures within the algorithm. This entails revealing the cause-and-effect relationships that account for how the algorithm generates its outputs, a pursuit that has roots in logic, philosophy of science, and computer science (Pearl, 2000; Spirtes et al., 2000). Another, not necessarily unrelated, way to open the black box is by explaining how a specific outcome was generated. This explanation may require providing a clear description of the steps, processes, or mechanisms that lead to a particular result, as discussed in numerous articles (Páez, 2019; Watson and Floridi 2021). More generally, making low-level mechanistic relations visible can be understood as conveying "useful information of any kind" about how the algorithm behaves and its outputs generated (Lipton, 2018).

On more practical grounds, there is the question of how can this latter form of transparency be achieved? To answer this question, we will refer to the classification provided by Guidotti et al. (Guidotti et al., 2019, 93:15), which outlines four methods for opening the black box: (i) explaining the model, (ii) explaining the outcome, (iii) inspecting the black box internally, and (iv) providing a transparent solution. Since methods (i) and (ii) have been covered in previous work (see Author), we shall exclude their analysis here. Method (iv), on the other hand, is closely related to (i), as it involves directly providing a model that is either locally or globally interpretable. We will not delve into the details of either of these methods but instead defer to the authors for their explanation (see (Guidotti et al., 2019, 93:14–15). We will, however, discuss method (iii), which focuses on inspecting the black box internally.

According to Guidotti et al., the process of inspecting a model involves providing a representation (which can be visual, textual, dynamic, static, etc.) that aids in our understanding of particular properties of the black box and leads to justification. For instance, sensitivity analysis plays a role in "observing the changes occurring in the predictions when varying the input [of the algorithm]" (Guidotti et al., 2019, 93:14). These changes can then be visualized, often through tools like partial dependence plots (Goldstein et al., 2015) and variable effect characteristic curves (Cortez and Embrechts 2013). The information extracted from various visualizations and plots contributes to the justification of the belief that the output has scientific value. Importantly, what distinguishes the process of inspecting a model is that sensitivity analysis focuses on analyzing specific properties of the black box without necessitating a comprehensive understanding of the entire system (Guidotti et al., 2019, 93:14).

A concrete example of inspecting a model is *Qualitative Input Influence* (QII). At its core, QII quantifies the joint influence that specific inputs have on the outputs of machine learning or Deep Neural Networks (DNNs). Datta, Sen, and Zick describe the fundamental principles of QII as follows: "A transparency query assesses the influence of an input on a quantity of interest, where the quantity of interest represents a system's behavior for a given input distribution" (Datta et al., 2016, p. 599). These assessments are later used to prepare *transparency reports* that accompany system decisions (e.g., explaining a specific credit decision) and for testing tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).

To illustrate how a transparency report works, consider the case of Mr. X, a 23-year-old adult male from Vietnam with an 11th-grade education, never married, with \$14k in capital gains and \$0k in capital loss (for a complete list of profile variables, see Fig. 4a in (Datta et al., 2016, p. 608). According to QII, Mr. X is classified as a low-income individual, despite having high capital gains and low capital losses. This output is somehow shocking, as "only 2.1% of people with capital gains higher than \$10k are reported as low-income" (Datta et al., 2016, p. 608). Given these unexpected results, there is a need to account for how this output is determined.

The transparency report can swiftly reveal which variables wield more influence over the output, thus justifying the belief that the result has scientific value. For instance, the report reveals that classifying Mr. X as a low-income individual is not due to his ethnicity or country of origin, as one might suspect without inspecting the algorithm. Instead, it's primarily attributed to his marital status, relationship, and education. This crucial insight is easily gleaned by examining the transparency report, which typically consists of a bar graph indicating the measured quantity for each variable (see Fig. 4b in (Datta et al., 2016, p. 611).

Admittedly, our description of QII is a simplified overview. A more comprehensive, though still incomplete, analysis would involve discussing various metrics used to measure the correlation between variables, the strength of these correlations (e.g., Pearson correlation), the weighting of protected attributes (e.g., race, gender, drug history, arrests), the proportion of positive predictions (e.g., Disparate Impact Ratio), the assessment of dependence between random variables (e.g., Mutual Information), and considerations of group disparity (i.e., classifiers that do not use variables as inputs (e.g., gender for a bank loan) that lead to group disparities tend to be fairer). Despite these simplifications, the essence of QII remains intact: the authors demonstrate how specific groups of variables (such as age, marital status, etc.) influence the machine learning model's output (e.g., Mr. X's classification as a low-income individual) through various measures made visible in the transparency report.

Let us close this section by noting that transparency encompasses a broader range of methods than our analysis of "opening the black box." We can attain forms of *functional transparency* without fully delving into the algorithm's inner workings or revealing its internal representations. This occurs when an output is explained in terms of the algorithm's high-level behavior. For instance, algorithms such as LIME can account for the predictions of any classifier by locally learning an interpretable model. In practice, if an ML system predicts that a patient has the flu, LIME can highlight the symptoms in the patient's history responsible for the prediction. 'Sneeze' and 'headache', for example, are key variables used by the algorithm. They are flagged as net contributors to the flu prediction. In contrast, 'no fatigue' is a variable used as evidence against the prediction (Ribeiro et al., 2016).

3.1.1 Objections to Transparency

In the pursuit of transparency, an array of resources has been dedicated to the cause. In this respect, it is imperative to recognize that transparency carries numerous shortcomings that are frequently overlooked. This section will briefly explore some of these issues and assess their potential impact on our confidence in the algorithm and on its outputs.

First, there is the issue of *algorithmic regress* or *transparency regress*, which becomes apparent when considering the fundamental goal of transparency –to unveil the internal mechanisms, causal connections, and interdependencies within an algorithm. In pursuit of this objective, researchers commonly employ an interpretable predictor (referred to as IP_1), designed to elucidate the generation of a specific output (Doshi-Velez and Kim 2017). However, the challenge arises when we realize that, in principle, there is no inherent reason to believe that IP_1 accurately represents the algorithm's inner workings. It is conceivable that IP_1 may harbor biases, oversight of key internal mechanisms, or instances of manipulation, such as reporting forms of "transparency" favoring specific groups' interests. The algorithm COMPAS could be transparent in ways that align with Northpointe interests; QII could produce transparency reports that favor the bank's interests. To address this issue, we need to somehow ensure the transparency of IP_1. The best way we know to do so is by means of another interpretable predictor (IP_2). Yet, this move only reintroduces the same concerns present earlier, perpetuating the cycle of transparency regress. In this context, there are no safeguards preventing us from suspecting that any interpretable predictor may possess faults or deficiencies.

Arguably, there are two ways to address transparency regression. Either we consider a primal interpretable predictor that is surveyable, contestable, auditable, and overall sanctionable by humans (designated as IP_n);[7] or we take a leap of faith and accept any given interpretable predictor as reliable. In the former case, regressing down to a primal interpretable predictor is pragmatically undesirable since the accumulation of IPs makes the entire enterprise of opening the black box utterly useless. In the latter case, there is an epistemic pressure to provide reasons as to why an algorithm is reliable when our means of justification (i.e., transparency) are unconvincing.

Another challenge for advocates of transparency, related to transparency regress, is the need to demonstrate that the transparency of any IP_n entails the transparency of IP_{n-1}, which in turn entails the transparency of IP_{n-2}, and so forth. That is, we need to show that the succession $IP_n \rightarrow IP_{n-1} \rightarrow IP_{n-2} \rightarrow \ldots \rightarrow IP_2 \rightarrow IP_1$ effectively maintains justification. In principle, transparency is possible, but in practice, it either involves a pragmatically undesirable transparency regress or a — possibly ungrounded — commitment to an arbitrarily chosen interpretable predictor.

The second objection is that reliance demands a sense of cognitive security that transparency might not be able to provide. The primary issue is that, for a transparent algorithm to be considered reliable, we must not only reveal the inner workings of the algorithm but also demonstrate a comprehensive understanding of it. For example, demonstrating that a mole is classified as melanoma based on specific conditions (e.g., a size larger than 6 mm, asymmetrical, etc.) does not guarantee that we understand why this classification occurs or even that it is the correct classification. To illustrate this point further, consider again the Convolutional Neural Network (CNN) that analyzes ID photos of individuals, identifies facial traits, and classifies each photo as either belonging to a 'criminal' or 'non-criminal' (Wu & Zhang, 2016; Wu & Zhang, 2017). While we can show how an algorithm produces a given classification, it is an overestimation to claim that we have understood the sources of criminality or that we have reasons to believe the output. Transparency seems to be able to provide, at best, the former but not the latter two.

3.2 Computational Reliabilism

Transparency posits a perspective that relies on surveying the inner workings of an algorithm to justify its outputs. As mentioned, the merits of this viewpoint encounter difficulties under certain conditions. This is not to suggest, of course, that we should abandon the pursuit of transparency. The value of transparency as an ideal is not in question. However, we must be cautious not to conflate our pursued goals with the legitimate ends of inquiry. The search for transparency oftentimes blurs the line between the valued and the valuable, and what is effectively feasible.

[7] We are not advocating for requiring all of those practices and properties of algorithms. However, it remains an open question which subset is sufficient for the purposes outlined here.

The alternative to transparency that also fosters reliance on algorithms is to embrace their black-box nature. In other words, our justification in believing certain outputs could no longer depend on opening the black box. What might seem like an acceptance of defeat is, in fact, a proposal for a new strategy for justifying our beliefs. Computational reliabilism (CR) was initially developed for computer simulations (Durán & Formanek, 2018; Durán forthcoming) and has recently been discussed in the context of medical AI (Durán & Jongsma 2021). The concept behind CR is simple and appealing: beliefs formed by reliable computationally related processes are better justified than those formed by unreliable ones. Advocates of CR argue that these beliefs do not necessarily arise from revealing the inner workings of the algorithm but from established practices, standards, methods, metrics, and a wealth of knowledge inherent in the design, development, use, and maintenance of algorithms. Importantly, none of these depend on employing a third-party algorithm (i.e., an interpretable predictor). Furthermore, CR operates under a dispositional theory that accepts occasional errors and misclassifications as long as, overall, the algorithm is reliable –that is, it produces outputs with scientific value. Formally, a reliable algorithm is defined as a belief-forming process that consistently renders outputs of scientific value more often than not. Under this heading, we must ask: what makes an artificial intelligence system reliable? According to CR, three token *reliability indicators* can be identified:

- *RI_1Technical performance of algorithms*: it focuses on the design, coding, execution, and other technical aspects of artificial intelligence systems that make the system robust, including the collection, curation, storage, and analysis of data;
- *RI_2Computer-based scientific practice*: focuses on the practices incumbent to ML-based scientific research and which results from the implementation of scientific theories, principles, and hypotheses, as well as the interactions, debates, and other ways of engaging in standard scientific research; and finally,
- *RI_3Social construction of scientific beliefs*: focuses on the broader goals of accepting the AI and its outputs in diverse communities (e.g., scientific, academic, general public, etc.) through debates and other forms of intellectual exchange.

Let us now briefly consider each reliability indicator in turn. Take RI_1, where reliability primarily arises from enhancing the robustness, precision, and accuracy of AI, thereby reducing the error rate. Verification and validation methods, encompassing various sub-categories (see, for instance, (Oberkampf & Roy, 2010), exemplify approaches aligned with this goal. Achieving high accuracy and minimizing errors indisputably enhances the reliability of algorithms. Of course, these methods vary among systems, since validation methods for computer simulation are, in important ways, different from machine learning (Boge, 2022). Consequently, the quality of an algorithm's outputs is not solely contingent on its numerical proximity to a 'ground truth.' Outputs also hinge on the user's comprehension of their scope, its suitability for the intended purpose, embedded assumptions, trade-offs made for tractability, and the algorithm's representative performance. Thus understood, RI_1 shifts the focus from the properties of algorithmic outputs (whether they are accurate or not) to the properties of the inquiry methods themselves (e.g., the appropriateness of verification and validation methods for specific goals). In this manner, high precision, accuracy, and a low error rate come with the same assumptions and considerations as the methods that bring them about.

RI_2, on the other hand, directs attention to how scientific theories, hypotheses, principles, and other propositions grounded in science are operationalized into the algorithm or the databases used. It is noteworthy that such embedding may not always occur explicitly and intentionally. Researchers might not consciously operationalize a specific set of scientific propositions into the algorithm. AI systems, particularly when applied in fields like medicine, have the ability to distill scientific knowledge from extensive literature reviews, scientific debates, and various sources. Notably, machine learning and deep neural networks in medical applications often leverage this principle. Given the impracticality or undesirability of explicitly implementing a medical theory into the algorithm, medical machine learning is often trained by selecting and cohesively assembling medical knowledge drawn from reputable journals. An illustrative example is Benevolent AI, a machine learning-based system in drug discovery that asserts its ability to 'capture the interconnectivity of all relevant available data and scientific literature using their proprietary Knowledge Graph' (see https://www.benevolent.com/what-we-do).

RI_3 aims to capture the scientific debates conducted with AI methods, emphasizing active involvement rather than mere automation. In a typical scientific setting, algorithm outputs are subject to comprehensive scrutiny and testing within the relevant community before their acceptance. To illustrate this intricate process, consider discovering a new drug. Before it reaches the market for its intended purposes, it must traverse a series of rigorous stages, including clinical control testing, pilot studies, and scientific debates. This journey culminates in final approval for human use, requiring collaboration with other scientists. This collaboration involves engaging in debates on result interpretation, scope, limitations, and, wherever possible, replication. Furthermore, approval of a new drug also requires independent testing by authorized institutions, such as the FDA in the US and the EMA in the EU. These components collectively contribute to the justification of the output, ensuring that they withstand collective scrutiny and meet the highest standards before integration into practical applications. In this respect, commitments to reliable AI extend to a comprehensive network of scientific methodologies, standards, results, and established traditions. As aptly noted by Elgin, this network enables scientists to build upon each other's work with confidence, ensuring that justified outputs align with the epistemic value prescribed by their respective disciplines (Elgin, 1996, p. 77). Naturally, within this network, disputes and disagreements are expected, encompassing conflicts related to (moral, scientific, political) values, methodological approaches, and the operationalization of varying concepts, theories, and other units of scientific analysis.

Earlier, we referenced BenevolentAI in the context of reliability indicators RI_2. The subsequent debate following BenevolentAI's output, particularly the revelation of baricitinib as a promising candidate to combat COVID-19 effects (Medeiros, 2021), serves as an illustration of how the justification of beliefs can be strengthened through scientific disputes and controversies. Favalli and colleagues, reporting on potential harms associated with baricitinib administration, notably an increase in herpes zoster and herpes simplex infection in specific patient groups (Favalli et al., 2020), prompted a reevaluation of the drug's target patients. The team implementing BenevolentAI, in agreement with Favalli's concerns, exercised caution in recommending the drug for those patients (Richardson et al., 2020). Notably, this debate played a pivotal role in determining the requirements for justifying AI outputs, establishing which errors and artifacts are tolerable, and validating the soundness of underlying assumptions. In essence, it showcases the dynamic and evolving nature of the discourse surrounding AI, emphasizing the importance of rigorous examination and collective consideration in shaping the future trajectory of this field.

Finally, it is important to highlight that CR represents a return to established scientific methodologies and practices, albeit with a unique twist. Now, researchers are compelled to integrate well-accepted principles of algorithmic design, utilization, and maintenance. According to CR, this integration enhances researchers' confidence in AI systems, justifying their belief in the scientific merit of the outputs, and ultimately fostering the reliability of AI. Remarkably, all of this is achieved without opening the black box.

3.2.1 Objections to Computational Reliabilism

Just as we observed with transparency, CR also has important challenges to overcome. A notable concern arises from the frequency at which beliefs are justified. While in many instances, the algorithm's output may indeed have scientific merit, leading researchers to deem the system reliable, there's a valid worry that the rare instances of system failure could have profound implications. To illustrate this, consider a medical AI providing various oncological diagnoses. Assume the system is generally deemed reliable because its outputs align with diagnoses made by human oncologists, demonstrating scientific merit. Users trust and treat the medical AI accordingly. Now, envision a scenario where the system misdiagnoses one single patient, inaccurately categorizing them as healthy instead of detecting a form of cancer. Under CR, even if this specific output lacks scientific merit, the medical AI as *a whole system* is still considered reliable. The critical question that arises is whether physicians are epistemically entitled to rely on the system after such a failure or if a significant reevaluation of the conditions under which the system operates is imperative. This example underscores the potential limitations and challenges associated with relying on CR in complex, high-stakes domains such as medical diagnosis.

A second limitation of CR is associated with the availability of reliability indicators. It is improbable that we are in possession of all the pertinent indicators for a given AI system.[8] In such scenarios, researchers are tasked with evaluating the reliability of their system based on a limited set of indicators. Furthermore, the few

[8] Equally crucial is to acknowledge that not all reliability indicators are universally applicable. For instance, while validation might be more pertinent for empirically-driven AI (e.g., climate change and mental mechanisms), it might hold less relevance for theoretically-driven AI (e.g., the origins of the universe and protein folding).

available indicators may wield disproportionate influence over the attributed reliability of any AI system. To illustrate this counterfactually, our assessment of the reliability of a system would most likely differ had we had access to all the relevant indicators. We term this phenomenon the *tyranny of the few*, underscoring the importance of having available as many and as diverse reliability indicators as possible. Ultimately, it is still unclear how many reliability indicators are necessary to mitigate the tyranny of the few.

Despite these concerns, CR represents a significant advantage in evaluating the reliance on AI systems. One key aspect is the "decentralized" nature of the reliability indicators. This means that there are various sources of indicators available to us and that these sources operate independently from each other (e.g., validation is not contingent on scientific debates). Another crucial advantage of CR is that humans *are* in the loop in a meaningful way. This contrasts with transparency, where humans typically play a passive role in trying to understand an explanation or an interpretable predictor.

4 The "Extra Factor"

Now we turn our attention to discussions revolving around the "extra factor". We present two main positions in the specialized literature in connection with the conceptual and normative possibility of trusting (or not trusting) AI systems. It is important to note that these positions are rather absolute in their views and in direct opposition to each other. Whereas one states that trusting AI is either not possible or undesirable, the other advances claims for its plausibility. In what follows, we discuss each one in turn.

4.1 Trust in AI is neither Possible nor Desirable

In section 2, we mentioned how interpersonal accounts of trust place humans at the center of their analysis. Drawing on similar philosophical ideas, adversaries of the possibility of trusting AI base their skepticism on the (rather obvious) differences between humans and machines. In this context, two main claims are set out. The first claim is that trust in AI is conceptually impossible because genuine trust in an inanimate entity (such as AI) is a category mistake. Scholars endorsing this claim typically argue that trust in AI would be incompatible with any philosophical account of interpersonal trust. The second claim is normative in nature and states that we *should not* place our trust in AI systems since this would lead to undesirable consequences. These amount, for instance, to the fact that a responsibility gap emerges given that trusting an AI system enables AI developers and designers to elude their (moral) responsibility by outsourcing it to AI systems (Starke et al., 2022). Of course, these two claims are not to be considered completely separated: usually, authors who deny the theoretical possibility of trusting AI also endorse the claim that AI systems are entities that *should not* be trusted. However, we keep these claims separate for analytic purposes. In what follows, we critically discuss accounts supporting these two positions and point out that they represent a considerable challenge to anyone defending the possibility and desirability of trust in AI.

We could identify three main positions pertaining to the *impossibility* or *undesirability* of trusting AI systems. In the following, we address each one in turn. The first account is known as the *affective account of trust* and consists of identifying the "extra factor" with the favorable disposition or goodwill of the trustee to fulfill the particular goal they have been entrusted with. This requires that the "trustee is favourably moved by the trust placed in them" and that "the trustee has the trustor's interests at heart" (Ryan, 2020, p. 12). This account of trust emphasizes the value of the interpersonal aspects of the trust relationship, such as emotions, psychological states, and motivations (Ryan, 2020). For example, as Ryan points out, when we trust our friend, we assume – following the affective account – that she is willing to keep our secret because she does not want to wrong us and not because she would otherwise run into trouble. That is to say, the motivations behind her willingness to keep our secret come to the fore in the affective account of trust: our friend does not keep the secret merely for self-interest but rather because she cares about us. These considerations make clear the anthropocentric nature of the affective account of trust and the difficulty of successfully applying it to trust considerations in which the trustee is not human, as in the case of AI systems. In fact, it seems out of place to expect an AI system to have motivations and affective attitudes in the first place. As we will elaborate on later in this section, this is a central point on which arguments for the impossibility of trusting AI systems hinge.

Let us now turn to the second account of interpersonal trust. This is known as *normative trust* because it refers to the normative expectations that the trustor has on the trustee. This account takes that the trustee *ought to* fulfill the commitments that emerge when the trustor decides to entrust her with a certain goal or task.[9] For instance, if a friend asks us to keep her secret, we *should* do so in virtue of the fact that she is entrusting us with a piece of information that we are not supposed to share. Clearly, this account requires the trustee to be the bearer of moral responsibility. In particular, in case a breach of trust occurs, the trustee needs to be a suitable receiver of blameworthiness. As Hatherley points out, "I rely on you when I predict that you will behave in a certain way, though I trust you when I judge that you ought to behave in a certain way." (Hatherley, 2020, p. 3) It is obvious that both the affective and normative accounts require that the trustee is aware of the fact that the trustor has placed trust in them. Once again, these human-centered demands seem difficult to fulfill for AI systems as inanimate entities.

The third account, known as the *rational choice account*, sees the trustor as making a rational evaluation when deciding to trust the trustee based on the likelihood that the trustee will behave as intended towards the fulfillment of a certain goal. This does not entail any kind of demand (normative or otherwise) on the trustee for the trustor to engage in a trust relationship. It also does not require the trustee

[9] As Ryan rightly points out, this does not mean that the trustee will have to fulfill every task she has been entrusted with. The moral acceptability of the particular task in question needs to be secured before saying that the trust relationship entails normative expectations.

to be moved by the "right reasons" to act as the affective account postulates. It only requires that, based on a regular frequency, the trustee behaves as intended. So, contrary to the other two accounts of trust, motivations and normative expectations do not play a central role in the rational choice account. As such, rational trust "is reliant on specific features of a situation, rather than the relationship between the trustor and the trustee." (Ryan, 2020, p. 11).

Quite intuitively, affective and normative trust set a standard for the extra factor that cannot be fulfilled by AI systems qua inanimate entities without attributing to them genuinely human traits (e.g., agency, emotions, motives, etc.). In fact, if we were to consider affective trust for an AI system, we would need to ascribe to it some forms of human agency and emotional states (awareness, empathy, compassion) to be able to say that the system is "willing" to live up the demands of a trust relationship. However, attributing these genuinely human traits to AI systems seems to be unwarranted. Moreover, and as mentioned before, it seems inappropriate to have sentiments of betrayal and deception — that usually would be in place when affective trust is broken — towards inanimate entities.

One can argue along similar lines when it comes to the normative expectations that are central to the normative account. Since AI systems are unaware of any form of trust that we may pose in their functioning, normative expectations on their performance (i.e., that they should work as we trust them to do) would be utterly misplaced. Therefore, due to the impossibility of AI systems being the appropriate bearers of normative demands, according to Hatherley, "the pursuit of trustworthy AI represents a notable conceptual misunderstanding" (Hatherley, 2020, p. 3). In the face of what has been said so far, the attribution of trust to AI systems would require us to anthropomorphize AI systems by attributing to them relevant human traits (such as some forms of agency or consider them receivers of moral responsibility, for instance).

In the face of these considerations, a rational choice account of trust seems better suited to sanction trust in AI systems because it does not require genuinely human attitudes and motivations to be in place. That is to say, the rational choice account offers a way to avoid the anthropomorphization entailed by the affective and normative accounts. This is the case because this account does not demand attention to the motivations of the trustee; the extra factor rather amounts to "a rational calculation of whether the trustee is someone that will uphold the trust placed in them" (Ryan, 2020, p. 4). However, the rational choice account also does not come without costs. In fact, some scholars consider the rational calculation of the likelihood that the trustee will perform as the trustor expects does not qualify as an 'extra factor' at all but rather boils down to mere reliance. According to Nickel et al., (2010) this comes to light because the rational choice account "is unfit to explain why performance-failure in cases of genuine trust leads to appropriate feelings of blame or betrayal in cases of malevolent breach of trust, and directed anger or disappointment in cases of negligence or incompetence" (p.431) This claim is also shared by Ryan (2020, 11). These authors' main concern is that the rational choice account does not allow us to make a conceptual distinction between morally-loaded reactions of betrayal occurring when trust is breached or being merely disappointed when someone or something we rely on fails to meet our expectations. However, as pointed out in the previous section, this distinction is pivotal to standard philosophical accounts of interpersonal trust (see, e.g., Hawley (2014).

The limitations pertaining to each account of trust when inanimate objects such as AI systems need to be accounted for work in support of positions arguing for the impossibility of trusting AI systems. This often leads authors to the conclusion that trustworthy AI is a conceptual mistake and “one needs to either change ‘trustworthy AI’ to ‘reliable AI’ or remove it altogether” (Ryan, 2020, p. 17). Thus, authors who hold a skeptical position regarding trustworthy AI take that AI systems cannot be seen as genuinely trustworthy because (1) it is either impossible to trust AI systems without anthropomorphizing them (see the critique to the affective and normative accounts), or (2) an account of trust that does not require AI systems’ anthropomorphization fails to maintain the conceptual distinction between morally-loaded trust and mere reliance, thus rendering the debate about trust in AI obsolete (see the critique to the rational-choice account of trust). Against this background, trusting an AI system would amount to misplaced trust (Ryan, 2020, p. 4).

Let us now turn to the second claim, that is, that trusting AI systems is *undesirable*. Starting from the assumption that trust needs to be a relation between peers in which beliefs and promises are made, Bryson (2018) defends both claims, i.e., that trust in AI *cannot* occur and *should not* be pursued. Bryson particularly emphasizes why we should refrain from ascribing to AI human-like features such as trustworthiness. The danger in doing so lies, according to Bryson, in the fact that developers and companies owning AI systems could use this to outsource their responsibility to these systems and evade moral blameworthiness when something goes awry. In Bryson’s words: “malicious actors will attempt to evade liability for the software systems they create by blaming the system’s characteristics, such as autonomy or consciousness.” (Bryson, 2018) In the face of these considerations, she concludes by stating that “AI is not a thing to be trusted. It is a set of software development techniques by which we should be increasing the trustworthiness of our institutions and ourselves.” (Bryson, 2018) Thus understood, the undesirability of attributing trust to AI systems amounts to the fact that, among others, a responsibility gap would emerge. Moreover, it would confer to AI systems capabilities that need to remain in the domain of human expertise, such as accountability and autonomy, creating unrealistic expectations of what AI systems can effectively achieve. A similar critical position is also shared by Tallant, (2019), who states that efforts pushing forward trust in automated cars, for example, are nothing else than a marketing move (Tallant, 2019, p. 116).

Along similar lines but focusing on the nature of trust in medical contexts, DeCamp & Tilburt (2019) advance the claim that talking about trust in AI could lead to a decrease of trust in medical practitioners since they could, on occasions, not achieve the level of accuracy secured by some AI systems. However, mistakenly confounding the reliance of AI systems’ performance with a proper, morally loaded notion of trust can lead to devaluing physicians’ abilities and expertise. As DeCamp and Tilburt point out “(p)romulgating trust in AI could erode a deeper, moral sense of trust.” And continues: “(t)rust properly understood involves human thoughts, motives, and actions that lie beyond technical, mechanical characteristics. To sacrifice these elements of trust corrupts our thinking and values.” According to these authors, therefore, we should not put our trust in AI systems if we want to preserve the importance of the morally loaded form of trust we are ready to put into human physicians (DeCamp & Tilburt, 2019).

To sum up, the main reasons advanced by authors critiquing the possibility and desirability of trusting AI can be boiled down to the following points. First, trust in AI would lead to the danger of impoverishing the notion of interpersonal trust in its morally loaded sense (Ryan, 2020), reducing it to not much more than mere reliance (see the critique of the rational choice account). This would blur the line between the two clearly distinguished concepts of reliance and trust. Moreover, it would lead to the impossibility of having a discourse about (genuine) trust in AI without falling into the trap of its unwarranted anthropomorphization (Ryan, 2020). In other words, neither the requirements for the normative nor the affective account can be fulfilled without attributing human traits to AI systems. Second, even under the assumption that is was possible, trust in AI is undesirable because it would lead to the unjustified attribution of responsibility to computational systems, representing a possibility for designers, developers, and companies to evade (moral) duties intrinsic to their professional role. This seems to be particularly unsatisfactory in situations in which the allocation of responsibility and blameworthiness plays a particularly salient role.

In the face of these substantial criticisms regarding the very conceptual possibility and normative acceptability of trusting AI systems, several efforts have been made to respond to these critiques. In the next section, we analyze different positions of scholars attempting to conceptualize trust (and trustworthiness) so that it can be meaningfully used in AI-mediated contexts. We will present the most prominent positions and critically analyze their merits and shortcomings in view of what has been discussed so far.

4.2 Trust in AI is Possible and Desirable

While the arguments supporting the idea that genuine trust in AI is not possible have merit, excluding the prospect of placing trust in AI systems might still feel unsatisfactory. In fact, given the ubiquitous nature of AI-based technologies, they play a relevant role in mediating interpersonal relationships, and their influence is increasingly interwoven in our social structures (Eschenbach & Warren 2021). Moreover, the crucial role of trust in accommodating complexity and the fact that we are increasingly vulnerable to AI technologies have motivated scholars to continue pursuing a suitable conceptualization of trust in AI (Lee & See, 2004; Chen, 2021). However, simply applying accounts of interpersonal trust to situations mediated by AI systems does not seem a viable solution in the face of the issues discussed in the previous section. As Nickel, Franssen, and Kroes point out, "(a)ny applicable notion of trustworthy technology would have to depart significantly from the full-blown notion of trustworthiness associated with interpersonal trust" (Nickel et al., 2010, 429). In the remainder of this section, we sketch out some of the most prominent accounts in favor of trusting AI.

We could recognize four overarching approaches according to how the challenges raised in the previous section (i.e., the problem of anthropomorphizing AI system following the affective and normative account and the conflation of trust with mere reliance following the rational choice account) are faced:[10] (1) approaches that admit some form of (minimal) agency in AI (under these fall Chen's account of *trust-responsiveness* and Lewis and Marsh's *functionalist view* on agency and intentionality)(Starke et al., 2022; Chen, 2021; Lewis & Marsh 2022) - *agency-based approaches to trust*; (2) approaches that deny the normativity and affective dimensions of trust in AI, thus taking a non-normative and non-affective position (Ferrario et al., 2020, 2021) – *the rational choice account of trust*; (3) approaches that take a normative stance but without making AI systems the bearer of moral obligations (Nickel, 2022) – *the discretionary view on trust* ; and, finally, (4) *reductive accounts of trust*[11] that take AI systems to be the indirect recipients of our trust (Sutrop, 2019).

Starke and colleagues are prominent advocates of the first approach mentioned, i.e., the agency-based approach. These authors build their "argument on the rather strong assumption that one can reasonably attribute agency to AIs." (Starke et al., 2022, p. 157) They do not ascribe full agency to AI systems (in the sense of mental states such as beliefs and desires) but rather a form of minimal agency or agency in a weak sense. Such a minimal agency stems from the embedding of AI systems in socio-technical contexts along with their per-designed ability to adapt, evolve, and influence it. To make their case, the authors consider Latour's case of the Berlin key that cannot be removed from the lock without locking the door (Latour, 2000). In this example, the key plays, by design, a role towards a certain goal, namely, making sure that the door is locked from the inside. In this context, the authors take that "by playing its part in a complex network of actors that would not be feasible without its material manifestation, the key contributes to the disciplinary relationship itself." (Starke et al., 2022, p. 157). For this reason, the key is not to be seen as a mere object but rather as an *agent* in a specifically described environment that contributes to the intended purpose. So, by analogy, if a key can be considered an agent in this minimal sense, these authors do not take it to be far-fetched to attribute agency to AI systems as well. Drawing on this assumption, Starke and colleagues take that trusting AI systems is possible if considered across three different dimensions. Those are intentionality, reliability, and competence. While reliability amounts to the avoidance of malfunctions and competence to validity and accuracy of predictions, the intentionality of a system can be perceived, again, along the lines of a weak sense of agency. Therefore, so goes the argument, if an AI system brings about discriminatory effects, one has less of a reason to trust its intentions (in the weak sense of the term). However, if, on the other hand, a system has a high level of interpretability, one has good reasons to trust the system's intentions to bring about a certain goal (Starke et al., 2022, p. 159).

A similar position that assumes some form of agency and intentionality in AI systems is also taken by Chen (2021) and Lewis & Marsh (2022). Chen assumes a form of "derived intentionality" in AI systems that stems from the ability to display what

[10] Of course, we do not have the pretense to be exhaustive in this regard.

[11] We adopt Nickel's label for this form of trust (Nickel, 2022, p. 5).

can be considered intelligent behavior, such as playing chess or performing natural language processing (Chen, 2021, 1435). Let us note that AI systems' intentionality is, also according to Chen, not to be understood in a strong sense. The author rather states that "(a)s products of human intentional action, they have a *prima facie* claim to some form of derived intentionality" (Chen, 2021, 1436). So, supporting a middle-way position between defenders of trust in AI and accounts that state only the occurrence of reliance without trust in AI, Chen sees what he calls *trust-responsiveness* as the most suitable alternative: "a disposition to prove reliable under the trust of others." (Chen, 2021, 1441). In order to achieve this, engineers need to put efforts into making sure that AI systems are reliable and robust so that we are justified to trust them (see Sect. 3).

In a similar vein, Lewis and Marsh take a *functionalist view* on trust, which focuses on how the system "functions, and how it is subsequently perceived and reasoned about by others" (Lewis & Marsh 2022, 44). According to these authors, excluding the possibility that agency and intentionality can be meaningfully attributed to AI systems would be an unjustified instance of human exceptionalism (Lewis & Marsh 2022, 47). That is to say, they take that it is not warranted to assume that intentionality, agency, and trust pertain exclusively to humans. On the contrary, from a functionalist angle, the ability of AI systems to betray our trust is deeply connected with their purpose and the possibility of deception. As such, considerations regarding the transparency of AI systems' goals, for instance, come to the fore when questions of whether we are justified in trusting them need to be considered (Lewis & Marsh 2022, 45).

Whereas these arguments support some form of agency in AI systems for the attribution of trust, one could still be skeptical that this is the right kind of agency for genuinely trusting AI. In fact, one could make the case that this form of minimal agency is not enough to consider AI systems to be able to live up to the normative expectations that characterize trust relationships. For instance, questions about the attribution of responsibility and accountability to these systems arise. How minimal is this "minimal agency"? Are AI systems to be considered as equally responsible and accountable as human agents? In sum, requirements in terms of minimal agency still raise concerns about the actual normative expectations of AI systems.

This brings us to the second main position on trust in AI, which does not require any form of agency for AI systems but rather focuses on advancing a non-normative and non-affective account. This position – i.e., the rational choice account of trust – is advanced by Ferrario and colleagues (Ferrario et al., 2020, 2021). Key to these authors' account is that trust in AI comes in degrees, and we do not need to consider AI systems as suitable bearers of affective or normative expectations in order to meaningfully say that we trust them. Let us, in particular, consider two of the three forms of trust conceptualized in their incremental model of trust. According to these authors, a minimal form of trust (they call it *simple trust*) is secured if we rely on a system without constantly updating our beliefs regarding its reliance. In their words, "trust involves economising on monitoring" (Ferrario et al., 2021, p. 437). That is to say, the readiness of the trustor to rely on the AI without control is the step needed to move beyond mere reliance and trusting an AI system. Consider, for example, a medical AI system that provides physicians with treatment recommendations for their patients. One first phase of reliance *only* is in place when the physician interacts with the system and forms beliefs regarding its performance, accuracy etc. In this phase, the physician engages in the evaluation of the system's performance to assess its reliability (as discussed in Sect. 3, this can be done in different ways). After a certain amount of positive interactions with the system, the physician will likely consider it reliable. At this point, she can start to rely on it without seeking further evidence supporting the fact that her reliance is indeed justified. As soon as the need to monitor the system disappears altogether and the physician is ready to rely on the AI without control, an instance of simple trust occurs (Ferrario et al., 2021). Simple trust is thus a property of the trustor (the physician) and not of the trustee (the AI system providing treatment recommendations) (Ferrario et al., 2020). Therefore, it is important to consider that for simple trust to be in place, we are not required to deem the system trustworthy overall. On the other hand, a situation in which we are ready to rely on an AI system without monitoring it and, on top of this, consider the AI to be trustworthy is, according to Ferrario and colleagues, the most complete form of trust in AI. They call this form of trust *paradigmatic trust*. The authors emphasize that the latter is what is usually referred to in the literature revolving around trustworthy AI, even though fulfilling the conditions needed to attribute simple trust would be enough to meaningfully talk about trust in AI. Not considering the affective and normative dimensions of trust in AI in their account of simple trust, these authors aim to maintain a conceptual distinction between trust and reliance without running the risk of anthropomorphization addressed in the previous section.

An objection to the notion of simple trust could still be advanced by questioning whether it substantially differs from mere reliance, as the authors claim. In fact, backed into their concept is the assumption that assessing the system's reliability requires constant monitoring that is no longer needed once its reliability is effectively confirmed. From that point on, we have simple trust in the system. However, it remains unclear why, after securing the system's reliability and giving up our critical monitoring, we go a step beyond relying on it. In other words, why does reliance need constant monitoring while trust does not? It seems plausible to think of situations in which relying on the fact that something will be the case does not require a constant update of our beliefs. Once again, questions emerge when normative (and/or affective) considerations remain unconsidered, and we still want to maintain a conceptual distinction between reliance and trust. In fact, for some authors, we cannot simply disregard the normative dimension of trust, excluding it from the picture. These considerations bring us to the next position on trust in AI.

The third position we analyze is advanced by Nickel (2022), who develops a *discretionary account of trust* in which the normativity of trust comes to the fore.[12] According to this view, trust manifests in the discretionary authority that, for

[12] While the normative dimension of trust is central to the discretionary account, it does not encompass an understanding of trust in affective terms. So, while it takes a very different stand regarding the normativity of trust in AI compared to the rational choice account, it shares with the latter a lack of emphasis on motives and desires. We thank an anonymous reviewer for encouraging us to clarify this point.

example, physicians decide to attribute to a medical AI involved in medical decision-making. Discretion is understood as a "circumscribed authority accorded to another entity" (Nickel, 2022, p. 7) and, according to Nickel, "(t)ransferring discretionary authority to another entity carries distinctive moral weight." (Nickel, 2022, p. 4). In the author's view, discretionary authority amounts to trust only if predictive and normative expectations on the trustee (i.e., in our case of interest, on AI systems) are in place. For example, consider a physician who decides to attribute discretionary authority to an AI system that estimates patients' likelihood of being admitted to the intensive care unit after surgery. In attributing discretionary authority to this system, the physician holds normative expectations on it as she expects the system to function as intended, i.e., to function as it *ought* to. The normative dimension goes, thus, hand in hand with the purpose and goal of the system, that is to say, with what the system has been designed and implemented for. As Nickel points out, "(w)hen such function-based expectations are relevant to the needs and goals of clinicians, they provide the basis for giving some of the clinician's own discretionary authority to the AI application, allowing it to (help) answer questions that previously went unanswered, or that were previously answered using other means." (Nickel, 2022, p. 7) In view of this, how can this normative dimension be accounted for without falling into the unjustified anthropomorphization of AI we discussed in the previous session? According to Nickel, when discretionary authority is attributed to an AI system, the AI is the *object* of a moral obligation but not its bearer. This means that physicians do not trust the AI system directly, they rather trust AI designers, developers etc. "*through* the AI application" (Nickel, 2022, p. 6). So, AI practitioners have the (moral) obligation to ensure that the AI system that has been granted discretionary authority functions as intended in respect of shared values such as fairness and efficiency, for instance (Nickel, 2022, p. 4). The discretionary account allows thus to preserve the normative dimension of trust without having to take a stand regarding the thorny issue of having to attribute some form of responsibility directly to AI systems. In fact, AI practitioners are taken to be responsible for guaranteeing that a certain AI system is up to the expectations of physicians who are ready to confer discretionary authority to it.

In view of what has been said so far, a possible limitation of Nickel's account is a lack of clarity about the actual locus of our trust. In fact, Nickel's view on trust can be objected to, as it seems to effectively amount to trust in AI practitioners since they are ultimately responsible for problematic outcomes brought about by an AI system. While Nickel leaves this question open in his conceptualization of trust in AI, there are authors (Sutrop, 2019) who clearly defend the position that the only possible form of trust in AI is in the human beings involved in the development of AI systems and not the systems directly.

The fourth and last position that we consider concerning trust in AI - the reductive account - has been defended, among others, by Sutrop (2019). The author argues that "when we speak about trust in AI, in reality, we are speaking about trust or distrust of individuals and institutions who are responsible for developing, deploying and using AI " (Sutrop, 2019, p. 512). This position thus takes that while AI systems can be meaningfully relied upon, the object of our trust can only be the humans involved in the development of AI systems (e.g., designers, engineers, computer scientists etc.). So, according to this account, we do not trust AI systems; our trust rather lies *exclusively* in the human beings behind their development, and they have the moral obligation to make sure that AI systems meet the expectations we pose in their functioning. However appealing, this position does not come without problems. For instance, the self-learning and adaptive abilities of most AI systems are an indication that it is not always clear to what extent computer scientists and engineers can foresee a problematic behavior of the system that is possibly perceived as a breach of trust by the end-user (say, a medical AI that leads to a misdiagnosis).[13] Therefore, it is not a straightforward solution to consider trust in AI as amounting to trust in the humans behind the development of the system instead of the system itself.

Considering what has been said so far, it becomes clear that views regarding what trust in AI amounts to strongly diverge. The responses to the critiques advanced by scholars who are skeptical regarding its conceptual possibility (and normative stand) are formulated in very different ways – all with their weaknesses and strengths. What is common to all the positions defending the possibility of trust in AI, however, is that some form of system reliability must be accepted.

5 Concluding Remarks

In this article, we put forward an analysis of trust and trustworthy AI aiming at dissecting its main components, possibilities, and limitations. Thus, the main aim of this article was to shed light on the nuances of a concept that is currently overly used in the literature to refer to a number of often vague and high-level ethical demands that AI systems need to satisfy. To this purpose, we started by making an analytic distinction between reliance and the "extra factor". Both requirements are present in standard accounts of interpersonal trust in the philosophical literature. With this distinction in mind, we considered two opposing views on how to secure the reliance of AI algorithms, namely *transparency* and *computational reliabilism*. We showed that even though some form of reliance is often taken for granted in the literature on trustworthy AI, it is not a trivial matter to find a way to successfully account for this necessary desideratum. We argued that transparency, understood as methods aiming at opening the black box, is typically taken to be the gold standard to assess the scientific merit of an AI system's output. Even though the search for transparency can be seen as intrinsically valuable, we argued that it suffers from considerable shortcomings worth debating. In this respect, two chief problems were presented and briefly discussed. One that shows that transparency might imply some form of *transparency regress*, in which case the justificatory status of the algorithm is pragmatically and epistemically compromised. The second issue is that transparency demands a kind of cognitive security difficult to obtain, thus casting doubts on the kind of understanding that it is able to offer. We also discussed computational reliabilism as the chief

[13] Since the discretionary account previously addressed also sees humans behind the development of AI systems as bearers of moral responsibility for the AI systems' outcomes, this objection can be also advanced to that account of trust in AI.

contender to transparency. Contrary to the latter, computational reliabilism does not require 'opening' black box algorithms. Instead, justification comes from *reliability indicators*, many of which are quite familiar to us as they draw from standard scientific practice. As we discussed, computational reliabilism is also limited in important ways. We mentioned the frequency by which beliefs are justified and the tyranny of the few. Despite these, computational reliabilism still proves to be a suitable method to account for the reliance condition needed to secure trust in AI systems.

Overall, the discussion in this part of the paper highlights the importance of epistemological considerations underpinning ethical concerns surrounding trust and trustworthiness in AI. While reliance is often tacitly and vaguely assumed to correlate with a system's accuracy, our analysis reveals a more nuanced and detailed picture. Crucially, our arguments underscore the necessity of directing research efforts toward ensuring that the reliability of AI systems is thoroughly addressed before engaging with the question of whether morally-loaded trust in AI is possible and, if so, what forms it should take.

With these results in place, we moved on to the analysis of the 'extra factor,' understood as the second component in the definition of trust in AI. Here, we showed that scholars holding a skeptical view regarding the very possibility and desirability of trust in AI systems advance convincing arguments that need to be accounted for. In particular, we discussed the unwarranted anthropomorphization of AI systems and possibly undesirable consequences in terms of responsibility gaps. As we further considered, the accounts of authors trying to respond to the criticisms advanced are many and contrasting. Among others, we sketched some approaches that try to exclude the normative dimension of trust from the picture, while others attribute to AI systems either some form of minimal agency or require to trace trust back to the human beings behind its development. Even though these efforts have merit, we pointed out that some issues remain unresolved. For example, it is unclear whether the agency attributed to AI systems accounts for the 'extra factor' or whether it is legitimate to exclude the normativity of trust and thus blur the line between (mere) reliance and genuine (morally robust) trust. As we have shown, the debate is vast, and opposing views are defended. This article reconstructs key fundamental aspects of the debate in an attempt to bring clarity and order to an otherwise fragmented debate.

Admittedly, the way in which we structured the debate necessarily leaves out important considerations about trust and trustworthy AI that deserve further attention. One of particular importance is the right level of stakeholders to consider. In this article, we narrowed down the scope of our analysis to individual interactions with machine learning systems. For instance, we referred to the trust (or distrust) that a physician can have towards an AI system providing treatment recommendations for their patients. However, this is not the only dimension across which trust can be established. As one can distinguish different levels on which information is created and transmitted,[14] along similar lines, one could say that trust relations

[14] Alvin Goldman (Goldman, 2019) pointed out that the creation and transmission of knowledge can occur throughout three different dimensions: interpersonal, collective, and institutional. We think that this consideration can be transferred also on how trust is established. For more on trust in institutions, see Alfano and Huijts (Alfano and Huijts, 2019).

develop across different dimensions such as interpersonal (or individual), collective, and institutional. While in this paper, we focused on the interpersonal dimension of trust in AI systems, further research is needed to spell out in clear terms how AI systems impact attitudes of trust at the collective and institutional levels in AI-mediated contexts.

Acknowledgements We would like Sara Blanco Peña for her insightful comments on an earlier draft of this paper and the two reviewers who pointed out key issues that needed to be addressed.

Author Contributions All authors contributed equally to the article.

Funding This work was supported by "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement 871042).

Declarations

Ethical Approval This article does not involve or require ethical approval or informed consent.

Consent to Publish No further consents are required except the one that derives from both authors to publish the manuscript should it be approved for these purposes.

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

References

Alfano, M., & Huijts, N. (2019). Trust in Institutions and Governance. *The Routledge Handbook of Trust and Philosophy* (pp. 256–270). Routledge.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Bryson, J. (2018). No one should trust AI. *AI and Global Governance*. https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html. Accessed 27 Jan 2025.

Bugel, S. (2023). Fake doctor who worked in NHS for 20 years found guilty of fraud. *The Guardian*. https://www.theguardian.com/uk-news/2023/feb/15/fake-doctor-zholia-alemi-nhs-guilty-fraud. Accessed 27 Jan 2025.

Burrell, J. (2016). How the machine 'Thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Chen, M. (2021). Trust and Trust-Engineering in Artificial Intelligence Research: Theory and Praxis. *Philosophy & Technology*, *34*(4), 1429–1447.

Cho, J. H. (2019). StRAM: Measuring the trustworthiness of computer-based systems. In: ACM Computing Surveys 51.6. issn: 15577341. https://doi.org/10.1145/3277666.

Choung, H., David P., & Ross A. (2022). Trust in AI and its role in the Acceptance of AI technologies. *International Journal of Human–Computer Interaction*, 1–13. https://doi.org/10.1080/10447318.2022.2050543

Cortez, P., & Mark J. E. (2013). Using sensitivity analysis and visualization techniques to Open Black Box Data Mining models. *Information Sciences*, *225*, 1–17.

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

Datta, A., Sen, S., & Zick Y. (2016) Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. IEEE. https://doi.org/10.1109/SP.2016.42

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

Doshi-Velez, F., & Kim ,B. (2017). Toward*s a Rigorous Scienc*e of interpretable machine learning. https://arxiv.org/abs/1702.08608. Accessed 27 Jan 2025.

Durán, J. M. & Formanek N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines, 28*(4), 645–666.

Durán, J. M., & Jongsma K. R. (2021). Who is afraid of Black Box algorithms? On the epistemological and ethical basis of Trust in Medical AI. *Journal of Medical Ethics*, *47*(5), 329–335.

Durán, J. M. forthcoming. Beyond transparency: computational reliabilism as an externalist epistemology for algorithms. In J. M. Durán, & G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library: Springer.

Elgin, C. Z. (1996). *Considered judgement*. Princeton University Press.

European Commission. (2019). *High-level Expert Group on Artificial Intelligence*. Ethics Guidelines for Trustworthy AI.

Favalli, E. G., Martina-Biggioggero, G., & Maioli, and Roberto Caporali (2020). Baricitinib for COVID-19: A suitable treatment? *The Lancet Infectious Diseases*, *20*(9), 1012–1013.

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A Multi-layer Model of Trust to Analyze Human-Artificial Intelligence interactions. *Philosophy & Technology*, *33*(3), 523–539.

Ferrario, A., Loi, M., & Viganò, E. (2021). Trust does not need to be human: It is possible to Trust Medical AI. *Journal of Medical Ethics*, *47*(6), 437–438.

Fricker, M. (2007). *Epistemic injustice: Power and the Ethics of Knowing*. Oxford University Press.

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

Goldman, A. I. (2019). The what, why, and how of Social Epistemology. *The Routledge Handbook of Social Epistemology* (pp. 10–20). Routledge.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the Black Box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining Black Box models. *ACM Computing Surveys*, *51*(5), 1–42.

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. (2015). Trust and Distrust between patient and doctor. *Journal of Evaluation in Clinical Practice*, *21*(5), 798–801.

Hawley, K. J. (2017). Trust, distrust and epistemic injustice. In I. J. Kidd, J. Medina, & G. Pohlhaus (Eds.), *The Routledge Handbook of Epistemic Injustice* (pp. 69–78). Routledge.

Hawley, K. (2019). *How to be trustworthy*. Oxford University Press.

Humphreys, P. W. (2009). The philosophical novelty of Computer Simulation methods. *Synthese*, *169*(3), 615–626.

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

Kaur, D., Uslu, S., Rittichier, K. J., Durresi, A. (2022). Trustworthy Artificial Intelligence: A review. *ACM Computing Surveys (CSUR)*, *55*(2), 1–38.

Latour, B. (2000). The Berlin Key or how to do words with things. *Matter Materiality and Modern Culture*, *10*, 21.

Lee, J. D., & See K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lewis, P. R., & Marsh, S. (2022). What is it like to Trust a Rock? A Functionalist Perspective on Trust and Trustworthiness in Artificial Intelligence. *Cognitive Systems Research, 72*, 33–49.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., & Yi, J., & Zhou B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46.

Lipton, Z. C. (2018). The mythos of Model Interpretability: In machine learning, the Concept of Interpretability is both important and Slippery. *Queue*, *16*(3), 31–57.

McLeod, C. (2021). Trust. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. https://plato.stanford.edu/archives/fall2021/entries/trust/. Accessed 27 Jan 2025.

Medeiros, J. (2021). How tech is changing healthcare. From rapid development and rollout of the Covid-19 vaccines to the science of isolation, machine-learning-enabled gene editing and digitised medicine. *Wired*. https://www.wired.co.uk/article/future-health-trends. Accessed 27 Jan 2025.

Nickel, P. J. (2009). Trust, Staking, and expectations. *Journal for the Theory of Social Behaviour*, *39*(3), 345–362.

Nickel, P. J. (2022). Trust in Medical Artificial Intelligence: A discretionary account. *Ethics and Information Technology*, *24*(7). https://doi.org/10.1007/s10676-022-09630-5

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of Trustworthy Technology? *Knowledge Technology & Policy*, *23*(3), 429–444.

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Páez, A. (2019). The pragmatic turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pearl, J. (2000). *Causality: Models, reasoning and inference* (Vol. 110). Cambridge University Press. https://doi.org/10.2307/3182612

Phillips, A. (2017). 'They're rapists.' President Trump's campaign launch Speech two years later, annotated. *The Washington Post*. https://www.washingtonpost.com/news/the-fix/wp/2017/06/16/theyre-rapists-presidents-trump-campaign-launch-speech-two-years-later-annotated/. Accessed 15 Jan 2024.

Pozzi, G. (2023). Testimonial injustice in medical machine learning. *Journal of Medical Ethics*, *49*(8), 536–540.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44.

Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A., Rawling, M., & Savory, E., & Stebbing, J. (2020). Baricitinib as potential treatment for 2019-nCoV Acute Respiratory Disease. *Lancet (London England)*, *395*(10223), e30.

Ryan, M. (2020). In AI we trust: Ethics, Artificial Intelligence, and reliability. *Science and Engineering Ethics*, *26*, 2749–2767.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.

Starke, G., Rik, van den Brule, Elger B. S., Haselager P. (2022). Intentional Machines: A Defence of Trust in Medical Artificial Intelligence. *Bioethics* 36 (2): 154–61.

Sutrop, M. (2019). Should we trust Artificial Intelligence? *Trames: A Journal of the Humanities and Social Sciences*, *23*(4), 499–522.

Szalavitz, M. (2021). The pain was unbearable. So why did doctors turn her away? In Wired. Retrieved November 2024, from https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/.

Tallant, J. (2019). You can trust the ladder, but you shouldn't. *Theoria*, *85*(2), 102–118.

Von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607–1622.

Watson, D. S., & Floridi L. (2021). The explanation game: A formal Framework for interpretable machine learning. *Synthese*, *198*(10), 9211–9242. https://doi.org/10.1007/s11229-020-02629-9

Wexler, R. (2017). When a computer program keeps you in Jail: How computers are harming criminal justice. *New York Times*. https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html. Accessed 20 Jan 2024.

Wu, X., & Zhang X. (2016). Automated Inference on Criminality Using Face Images. *arXiv Preprint arXiv:1611.04135*, 4038–52. https://www.semanticscholar.org/paper/Automated-Inference-on-Criminality-using-Face-Wu-Zhang/1cd357b675a659413e8abf2eafad2a463272a85f

Wu, X., & Zhang X. (2017). Responses to critiques on machine learning of criminality perceptions (Addendum of arXiv:1611.04135). arXiv. https://doi.org/10.48550/arXiv.1611.04135